Department of Agricultural,
Food, and Resource Economics
MICHIGAN STATE UNIVERSITY

Staff Paper Series

Rural Labor and Long Recall Loss

Kate Ambler, Sylvan Herskowitz, and Mywish K. Maredia

December 2021

Staff Paper Series #2021-001

Rural Labor and Long Recall Loss

Kate Ambler
Research Fellow, Markets, Trade and Institutions Division (MTID)
International Food Policy Research Institute

Sylvan Herskowitz
Research Fellow, Markets, Trade and Institutions Division (MTID)
International Food Policy Research Institute

Mywish K. Maredia
Professor, Department of Agricultural, Food, and Resource Economics
Michigan State University

Abstract

Commonly used data collection practices use annual recall to capture individuals' labor activities over a year. However, long recall periods are likely to suffer from distortions and loss, particularly when work patterns are seasonal and informal. In a panel of rural households in Malawi, we use a survey experiment to test the effect of using long recall periods on the reported number of labor activities, labor supply, and types of work relative to those resulting from a set of shorter, quarterly interviews. We document large losses from the longer recall window, particularly on the extensive margin of labor supply with reductions of over 20%. These losses are greatest for periods furthest from the last survey round and are especially large among individuals whose labor supply is being reported for them, reaching as high as 50% losses for some outcomes. The composition of households' primary respondents, predominantly male and older, as well as differential effects by age both suggest that use of long recall may lead to meaningful biases by both age and gender in resulting data.

Acknowledgment

This project was funded by the CGIAR Research Program on Policies, Institutions, and Markets and was approved by the IFPRI Institutional Review Board and the Malawi Committee on Social Science and Humanities, National Commission for Science and Technology. We thank the MwAPATA Institute for giving us access to the baseline data collected under the Agricultural Transformation Initiative (ATI), GeoPoll for phone surveys, and Laura Leavens for research support.

1. Introduction and Background

Understanding peoples' productive activities over the course of the year is important both to document how people earn their livelihoods and also to ensure that poverty programs are well designed and well targeted. Measuring employment is especially challenging in settings where productive activities are informal, with irregular intensity of participation, and seasonal, where much of the effort and earning is concentrated in specific periods of the year. These characteristics are especially relevant in rural labor markets in low-income countries that rely heavily on employment in the agricultural sector.

In order to capture labor activities that may be largely seasonal and informal, standard data collection practices ask about the main and secondary productive activities of each household member over the past year. While existing evidence is mixed, longer recall windows may lead to losses in data and data quality.[1] One possible way to address this concern is to capture labor activities over a shorter recall period. In this paper, we use a survey experiment to compare reported labor participation using a long (annual) recall method to those resulting from a set of quarterly interviews that follow a similar structure but over shorter, ninety-day, recall windows.

A commonly used format for labor modules in multitopic household surveys asks the household's primary respondent to report on their labor contributions over the preceding year, typically inquiring about their primary and then secondary productive activities. While there is often a stated preference for having each household member respond for themselves, frequently, either when that respondent is unavailable or when survey resources are insufficient to allow for self-reporting of all members, the household's primary respondent will then report on behalf of other household members as well, "proxying" their responses.[2] To avoid total reliance on this long recall measure, surveys also frequently include questions about work activities over the past seven days. However, the extremely short duration of these weekly questions and their resulting lack of coverage and relevance to a full year of labor contributions in settings with highly seasonal work, limit their usefulness in using them to characterize longer windows of labor contributions.

---

[1] See de Weerdt, Gibson, and Beegle (2020) for an excellent review of the survey methods literature that includes an overview of both recall windows and proxy reporting.

[2] The Living Standards Measurement Study (LSMS) surveys follow a similar structure and has a similarly stated preference for self-reporting whenever possible. However, a review of six LSMS surveys by Desiere and Costa (2019) finds that the use of proxy reporting is still widespread, ranging between 24% in Nigeria up to 85% in Mali.

Recently, the global penetration of cell phone access has introduced phone surveys as a method of data collection that could allow for more frequent data collection with shorter recall periods at a lower cost. While weekly interviews have been shown to be less vulnerable to recall loss than longer windows (Heath et al. 2021), weekly interviews across an entire year or season would be prohibitively expensive for most research budgets and overly fatiguing for respondents. In our work we instead test a middle ground option, comparing quarterly (90-day recall) against a yearly recall question using a series of phone surveys.

We find that the long recall responses result in much lower recorded levels of any work involvement (22%), number of unique activities reported (24%), and number of months worked (20%) compared to those based on the quarterly measures, although we do not find significant losses in hours worked. These gaps increase as the time since the final interview grows. Proportional losses from long recall (relative to short recall averages) are between two and ten times larger depending on outcome variable for proxied individuals than for the household respondents who self-report their labor. Splitting the sample by self and proxy reports, we explore heterogeneity by gender and age of the respondent as well as of the household member. We do not see clear patterns of heterogeneity for self-reports or by the respondent's characteristics when reporting for others. We do, however, see that relatively younger household members have greater losses from long recall when their responses are proxied. Even in the absence of sharp heterogeneity by gender among either self-reports or proxies, the profile of primary respondents (generally household heads) means that women and younger household members will experience larger losses from long recall as a consequence of their greater likelihood of having their labor contributions reported by proxy.

This paper contributes to a growing body of work on the effects of recall periods in the survey methodology literature. A considerable body of evidence suggests that longer recall periods tend to undercount household consumption, provision of agricultural inputs, and negative health events relative to shorter recall windows.[3] The evidence on recall windows as it applies to labor activities, however, is more mixed. In a study based in Tanzania, Arthi et al. (2018) compare

[3] Das and Sanchez-Paramo (2012) find that one third of acute illness is unreported when comparing one month to one week recall periods. Other notable examples showing high sensitivity of consumption to recall windows include Beegle et al. (2012b), De Weerdt et al. (2016), Backiny-Yetna et al. (2017), and Di Maio and Fiala (2019). [3] Beegle et al. (2012a) find little evidence of distortions from longer recall lags in recording agricultural inputs and production.

reported contributions of farm labor in an agricultural plot-based module using weekly recall with those from an end of season recall window. Longer recall windows exaggerate the number of hours worked per person by a factor of four, although the reported number of active plots and individual level participation are undercounted. They also study in person versus phone surveys which modestly reduce the size of these differences. In our rural sample in nearby Malawi, we find similar patterns in a roster-based labor module; reductions in reported labor contributions using the longer recall periods relative to a shorter recall window, but no overall difference in hours worked. In urban labor markets, Heath et al. (2021) examine very short recall windows and find modest effects of switching from weekly to daily recall with higher reported self-employment spells but no impact on wage employment. However, Garlick et al. (2020) find that weekly versus monthly surveys did not influence data quality or reported microenterprise activities.

Telescoping, the crowding in of actions that, in reality, occurred just outside of the intended recall window, could lead to exaggeration of reported behaviors.[4] While this phenomenon can affect responses using either short or long recall, short recall estimates could be especially impacted when aggregating across multiple survey rounds in order to characterize longer windows of time. An additional consideration is that short recall periods require multiple interviews in order to cover the same range of time. The act of conducting repeated interviews could change responses or impact attrition (Arthi et al. 2018; Zwane et al. 2011). In our study, all individuals included in the analysis participated in the full set of four interviews so that any effects resulting from multiple interviews should be the same for both those who are randomly assigned to the long or short recall groups in the analysis.

This paper also links to a second area of the methods literature centered on the use of proxy responses for data collection. Of note, Bardasi et al (2011) find that males are more effected by proxy losses in reported agricultural labor activities than women, though this is reduced when the primary respondent is his wife or well-educated. By contrast, Serneels, Beegle, and Dillon (2017) find that proxying does not affect estimates of returns to education. The closest existing paper in this literature to our own is Kilic et al. (2020), which also intersects with the recall literature on labor measurement. In their work, they use two nationally representative surveys in Malawi, which

[4] Abate et al. (2020) provide an example of this in household consumption data.

were conducted in the same year but followed different research protocols, to show underreporting of certain labor activities in the survey using "business as usual" that allowed for proxy reporting and presence of other household members relative to another survey that required own response and privacy during the interview. Notably, they find that distortions from proxying are stronger when using a long recall window, one year, than a shorter one of just seven days. Our finding on heterogeneity by proxy status is in line with theirs, building on their work by showing potential biases in resulting data stemming from this interaction between proxy and recall windows.

Finally, the findings in our paper have implications for a range of labor linked papers, especially those in rural settings that rely on similarly constructed data.[5] Large, primarily descriptive literatures currently exist on rural labor, gender and age differences in labor contributions, and rural income diversification. Our results suggest that all of these estimates could be meaningfully impacted by reliance on long recall-based data and further distorted by the use of proxy reporting.

We discuss our survey experiment and data in Section 2 and empirical strategy in Section 3. Results are presented in Section 4, followed by robustness checks and discussion in Section 5, and conclusions in Section 6.

2. Data

Our survey experiment relies on a sample of households included in the MwAPATA Institute's Malawi Rural Agricultural Livelihood Survey (MRALS) conducted in the fourth quarter of 2019. The data is representative of farm households at the eight selected districts level—two districts in the Northern Region (Rumphi and Mzimba), four in the Central Region (Lilongwe Rural, Dowa, Kasungu, and Mchinji), and two in the Southern Region (Neno and Blantyre Rural) (Muyanga et al. 2020). This was a multitopic household survey conducted in person and collected information on demographics, health, socio economic status, time use, and extensive modules on agricultural production from a total sample of 3,259 households.

[5] See Dzanku (2020), Asfaw et al. (2019), Yeboah and Jayne (2018), Imai et al. (2015), Himanshu et al. (2013), Djurfeldt (2013), Haggblade et al. (2010), Ellis (1998), and Ellis and Freeman (2004) for some recent examples.

2.1 Sample and survey structure

From the MRALS base sample of 3,259 households only 2,435 households had contact phone numbers and were therefore eligible for inclusion in our phone-based survey. Stratifying by region, 1,505 eligible households were randomly selected to be included in our survey experiment.[6] In total, 1,020 households were successfully contacted and included in Round 1. Since our primary goal was to create a complete panel of households, in Rounds 2 and 3 we targeted only those who we had successfully reached in Round 1, interviewing 833 and 968 households respectively. In Round 4, all households were attempted again, yielding 1,281 interviews. 701 households were successfully reached in all four rounds of interviews and constitute the full panel sample used in the analysis. Descriptive statistics of our analytical sample are presented in Table 1, Panel A. We note that average household size in the panel sample is approximately five people, with little variation across samples. The average number of adults in the household is 2.4.

Interviews were conducted by phone. The primary respondent was the respondent from the initial, in-person baseline survey, who was typically the household head. The labor module asked this respondent about their own labor participation and additionally asked them to report the labor participation of up to two other adult household members age 18 to 65.[7] While survey protocols allowed interviews to be done with other household members if the respondent was unavailable, in practice this happened very infrequently. Panel B of Table 1 highlights the characteristics of the household heads. Seventy-five percent are men and they are 44 years old on average. Sixty-nine percent are in a monogamous marriage, and an additional 10% are in a polygamous marriage. These household heads had 6.6 years of education on average. Panel C shows the characteristics of the other household members in the survey.[8]

---

[6] We targeted only a subset of households due to financial considerations and expectations regarding sample size needed to effectively answer research questions.

[7] If more than two eligible adults were present, we randomly selected which two were included. When two additional eligible family members were eligible in the household, the respondent was asked about them in random order.

[8] There is slightly higher attrition among this group resulting from a requirement that individuals were living in the household in all four rounds in order to be included in the analysis sample, whereas the sample numbers for each round include all listed members.

Table 1: Summary statisics, attrition and balance

	Full sample	Round 1	Round 2	Round 3	Round 4	Panel	P-value module order
Panel A: Households							
Household size	5.017	5.060	5.050	5.104	5.002	5.073	0.993
Number of adults	2.428	2.422	2.439	2.436	2.428	2.454	0.857
Number of children	2.589	2.638	2.611	2.668	2.575	2.619	0.898
Sample size	1505	1020	833	968	1281	701	
Panel B: Household head (designated respondent)							
Female	0.250	0.250	0.251	0.241	0.257	0.247	0.324
Age	42.605	42.605	43.595	43.247	42.733	43.916	0.566
Married (monogamous)	0.687	0.687	0.697	0.699	0.681	0.693	0.469
Married (polygamous)	0.098	0.098	0.098	0.094	0.101	0.104	0.779
Years of education	6.488	6.488	6.603	6.700	6.513	6.649	0.663
Sample size	1505	1505	833	968	1281	701	
Panel C: Other adults (proxied respondents)							
Female	0.722	0.722	0.738	0.738	0.716	0.762	0.996
Age	29.419	29.419	30.263	30.035	29.375	30.418	0.544
Married (monogamous)	0.551	0.551	0.569	0.571	0.545	0.612	0.916
Married (polygamous)	0.071	0.071	0.070	0.069	0.070	0.079	0.648
Years of education	6.860	6.860	6.891	6.892	6.899	6.908	0.365
Sample size	1765	1765	926	1076	1499	466	

Notes: Summary statistics are calculated using the full sample collected in each round. The p-value is a test for equality in means between those assigned to ask the quarterly recall questions first or second in the final survey round.

Households were contacted approximately once per quarter over the course of a year. While the surveys were intended to be spaced evenly by three months between surveys, in practice logistical issues affected this timing and led to some variation in the actual interval between survey rounds. The dates and timing of each survey wave are illustrated in Figure 1. These quarterly surveys were designed to be brief, focused on labor activities of adult household members. The labor module followed a similar structure as that used in many household surveys where respondents are asked about their primary work activity over the preceding 90 days. Follow-up questions capture further details about that activity, which months they spent doing it, how much time they spent on it, and how much they earned. After reporting that activity, they were asked if there was a secondary activity that they were engaged in over that same 90-day window and, if so, to provide similar follow-up information.[9] After completing this, the respondent was then asked the same sequence of questions about other adult household members (up to two).

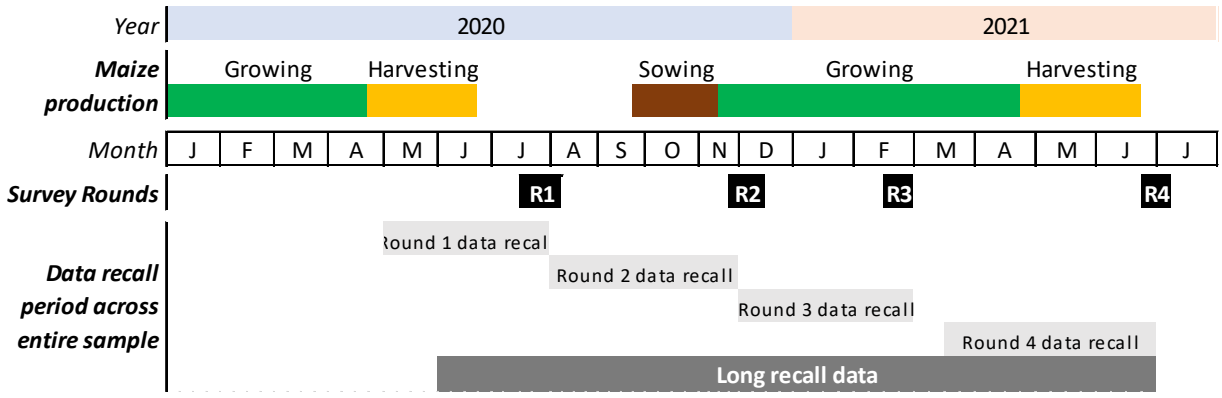


Figure 1: Maize crop calendar for Malawi, survey rounds, and data recall periods

In the final interview, an additional set of questions were included asking about individuals' primary and secondary activities over the past 12 months. In this final round, households were randomly assigned to either do the 90-day recall or annual recall questions first, for all family members. If annual recall were always reported after the short recall questions and the act of thinking, this could result in a systematic bias if the act of thinking about and reporting

[9] Respondents were then asked about their primary and secondary activities over the preceding week and, if distinct from the 90-day activities, details about those activities. Across all rounds, new jobs were reported in the 7 day category fewer than ten times.

labor contributions for the initial time frame affects the answers respondents provide on the latter set of questions. In the analysis, we use the reported labor participation of individuals using the recall window which they were asked first, to avoid the possibility of sequencing-induced bias. The analysis then leverages this randomization to test the effect of using long versus short recall windows in characterizing reported labor measures.

2.2 Outcome variables

Our analysis focuses on a set of four primary measures of labor supply. We examine two extensive margin measures: whether or not the person worked and the number of jobs they report working; and two intensive margin measures: the number of months the person worked and the numbers of hours worked.

We use the 90-day recall periods to construct quarterly estimates of each of these outcomes. For the intensive margin outcomes, we use the survey date and the variables indicating which months the indicated person worked to calculate the percentage of the month that is applicable for each individual. This is then used to calculate the months and hours worked in that 90-day period.[10] Using the 12-month recall data we create similar quarterly measures. We calculate these quarterly measures by using the question in the 12-month recall module that asks respondents to indicate in which months they worked in each job, and then use the survey dates from the earlier rounds to simulate the same time periods that are covered in an individual's 90-day recall questions.[11] Finally, for both the 90-day and 12-month questions, we sum the quarterly estimates to create equivalent aggregated measures. The main analysis compares the quarterly estimates and the aggregated long recall estimate.

In comparing labor aggregates using the short and long recall windows, two features of the data require attention. In the short recall estimates, respondents were able to list different jobs in each round so that, if they reported a different primary and secondary activity in each round, they could have up to eight unique activities.[12] Consistent with common survey practices, the long recall questions only allow for a maximum of two. The opportunity to report a greater diversity

---

[10] This is necessary because for a 90-day recall period, there are 4 calendar months in which the respondent may have worked. To estimate the months worked or hours worked, we must calculate the relevant fraction of the first and fourth month options.

[11] We use the day of survey for each household, so the period covered varies slightly by household based on the date of interview.

[12] We link jobs across survey rounds, so that if a respondent reports working the same job in all four rounds it is only counted as one unique job.

of work activities may be an advantage of the short recall estimates. A second difference is that when calculating hours worked we rely on a series of questions that ask respondents which months were worked in that time frame, the number of days in a typical month, and then the number of hours in a typical day. In the short recall responses, the number of days and hours apply to those months worked in that quarter, for the long recall those values are applied to all months worked throughout the year. The short recall responses therefore allow for greater variability in recorded labor supply within periods worked for the same activity.

Figure 1 provides a timeline of the four rounds of the surveys, showing the survey dates, and the months that are covered in each recall period, against the agricultural calendar for maize, the main staple crop in Malawi. Though in general the surveys were spaced quarterly, there was a longer delay between the third and fourth survey. This has the effect that some of the time covered in the 90-day recall asked in the Round 1 survey is not well covered by the 12-month recall questions asked at the endline. As such, we observe very low levels of work in Quarter 1 as measured in the 12-month recall data. In particular, we were concerned that including all four rounds of short recall from the four surveys would result in artificially low estimates of work in Quarter 1 for the long recall group. To address this issue, in our main estimates we do not include Quarter 1 and instead consider a "9 month" aggregate that comprises Quarters 2 through 4. The short and long recall data is more comparable using this approach and, as seen in Figure 1, those three quarters cover the main components of one maize growing season. Unfortunately, as a result, the data we use in the analysis does not quite cover a full 12 months, and there are some short gaps in coverage between survey rounds.[13] However, the three quarters of data still provide a meaningful period that we expect to cover the majority of individuals' labor activities.

3. Empirical Approach

The main comparison in our analysis tests for differences in individual-level labor outcomes as reported using the long-recall data, collected in Round 4, and those reported and calculated from the shorter recall data from the quarterly interviews. These outcomes are constructed as described

[13] Results using all four quarters show substantially larger aggregate differences between long and short recall, due to very low Quarter 1 estimates using the long recall.

in the previous section. We restrict our sample to individuals in households that were successfully contacted for all four rounds of the survey, and to individuals that were present in all four rounds. As discussed, because of concerns over the potential influence of asking 90-day recall questions just before annual recall questions in the final survey round (or vice versa) we randomized the sequence of these questions at the household level in the final interview. [14] Having randomized this sequence and to avoid biases induced by this ordering, we then drop the long recall observation of individuals who were asked about 90-day recall first, and we drop the short recall observation of individuals who were asked about annual recall first. We test for balance in that randomization and present the p-value for that test in the last column of Table 1, with no concerns noted. Because both groups were also interviewed in each four rounds, we can additionally be sure that the short recall group is not differentially affected by the repeated interviews and does not suffer from differential attrition.

Our main empirical specification, testing for differences between long and short recall is:

$$labormeasure_i = \beta_0 + \beta_1 longrecall_i + \boldsymbol{X}_i + \epsilon_i$$

The coefficient of interest is β_1, the difference in average reported labor measures between long and short recall. $\boldsymbol{X}_i$ is a set of gender by age group by reporting status (proxy or self) fixed effects. Age groups are defined as individuals under 25, 25-34, 35-49, and 50 and above, approximately in line with quartiles of the study sample. Robust standard errors are clustered at the household level.

In additional analyses we test for heterogeneity of these differences by proxy status, gender, and age. In these analyses we use fully saturated regression models to test each group's differences against a null hypothesis of no effect, but report p-values for differences across groups as well.

4. Results

4.1 Main outcomes

Table 2 presents the main results. In columns 1 and 2 we examine the extensive margin outcomes, whether the person worked at all and the total number of unique jobs reported. We find that the long recall led to large reductions in both of these measures. Long recall reduces any work participation by 20 percentage points relative to short recall, a 22% reduction. Just 8% of

[14] Appendix Table 1 shows that this ordering has a substantial impact on the long recall reported values in particular.

Table 2: Test of Losses from Long Recall Window

	(1) Worked at all	(2) Total activities	(3) Number of months worked	(4) Hours worked (100s)	(5) Worked hh farm/ag	(6) Worked non ag business	(7) Worked wage
Long Recall	-0.1987***	-0.2890***	-0.7191***	-0.1105	-0.2006***	-0.0615***	-0.0698***
	-0.0196	-0.0343	-0.1669	-0.2506	-0.0204	-0.0146	-0.0177
Mean Short Recall	0.9149	1.1897	3.6054	3.6601	0.8918	0.094	0.1365
Scaled Difference	-0.2172	-0.2429	-0.1994	-0.0302	-0.2249	-0.6541	-0.5114
Observations	1167	1167	1167	1167	1167	1167	1167
R-squared	0.3477	0.3141	0.2102	0.1494	0.3455	0.0382	0.0396

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.

individuals report no work whatsoever in the short recall, so an increase by 22 percentage points constitutes an enormous difference in reported labor force participation. Using long recall also reduces the number of unique jobs reported by 0.3, 24% of the short recall mean .

Turning to the intensive margin of labor supply, column 3 shows a similar pattern for number of months worked. Long recall reduces the reported number of months worked by 0.7 relative to short recall, 20% of the short recall mean. However, when considering hours worked, there is only a small negative coefficient that is not statistically different from zero. Though it is counter intuitive that no reduction in hours is measured given the reduction number of months worked, reporting for hours for the long recall is calculated using reported hours worked "in a typical month" and applying this number to all months worked across the whole year. The short recall estimates use a different reported measure of hours for each quarter. If respondents are inclined to report number of hours in a high intensity month, but this gets attributed to all months of participation, similar to behavior shown in Arthi et al. (2018), then this could lead to an exaggeration of activity in the long recall reported measures.

We also consider the types of work reported for each individual. In columns 5, 6, and 7 of Table 2 we report the impact of long recall on having worked on the household farm or in an agricultural-related home business, working in a non-agricultural related home business, or having done wage labor of any kind. Because the survey sample is exclusively rural, agricultural work is the most common, and 89% of the sample report having done it in the short recall group. There is a 22% reduction in that figure in the long recall group. These figures are comparable to those for having worked at all. Non-agricultural businesses and wage work are less common at 9% and 14% in the short recall group respectively. For these measures we see decreases in the long recall group of six percentage points for the non-agricultural businesses and seven percentage points for the wage work. Given low means, these translate into very large proportional effects of 65% and 51% suggesting disproportionate losses in reported non-agricultural labor contributions when using long recall that may be important sources of income diversification in this rural sample.

4.2 Heterogeneity by Self and Proxy Reporting

Next, we examine heterogeneous impacts of using long recall by proxy/self-reporting, gender, and age. First, we examine whether the labor statistic is self or proxy reported, presented in Table 3. Because a respondent is thought to have more complete knowledge of their own labor contributions than those of other household members, use of long recall when reporting on behalf

of others could be more taxing and more susceptible to omission. We estimate these effects by proxy status using a fully saturated regression model, such that the reported interaction is the long recall effect for each group. Beneath each column, we also report the mean in the short-recall group for self and proxy reports separately, the scaled differences, and the p-value for the difference between the two groups. In considering these patterns it should be emphasized that a household's primary respondent and therefore proxy or self-reporting status were not randomly assigned and therefore we cannot make causal claims of the effects of proxying on reported measures. We can show how absolute and relative gaps between long and short recall differ by proxy status. The survey respondents are almost exclusively household heads and, as reported in Table 1, are overwhelmingly male and significantly older than the proxied respondents.

Table 3 shows these results across the same outcomes as in Table 2. Across the first four outcomes losses from long recall are larger for proxied individuals, in both absolute and relative (to their mean) terms, although we lack sufficient precision to say whether these differences are statistically significant for the two intensive margin measures. Long recall loss for working at all among those who self-report is 5 percentage points, or 5% considering a short recall mean of almost 100%. The effect among those proxied is large: 42 percentage points or 54% of the short-recall mean for proxied individuals. These large differences between self and proxy reports may be due to increased recall bias in proxy reports or to the fact that household heads (who are the respondents in our survey) are engaged in more stable employment which is less likely to be forgotten regardless of reporting type. When considering total activities and months worked (columns 2 and 3), the proportional effect is larger than for working at all (15% compared to 5%) suggesting that recall bias may differentially affect the reporting of a second job or the specific months worked for those who are self-reporting. For these measures, the effect of long recall continues to be larger among the proxied respondents, both in absolute magnitude and percent effect. However, the difference is not statistically significant for months worked. As in the overall sample, there is no impact on hours worked, even while the point estimates are consistent with proxy reported labor being more vulnerable to long recall loss.[15]

[15] Regarding types of activities in columns 5-7 the patterns for farm work are similar to working at all. For non-agricultural businesses and wage work we observe larger absolute effects for those who are self-reporting; the proportional effect for non-agricultural businesses is larger for respondents and the effect on wage work is larger for proxies. In the latter case however, we cannot reject that the two effects are equal.

Table 3: Test of Losses from Long Recall Window by Self-Report or Proxy

	(1) Worked at all	(2) Total activities	(3) Number of months worked	(4) Hours worked (100s)	(5) Worked hh farm/ag	(6) Worked non ag business	(7) Worked wage
Long Recall X Self-Report	-0.0533***	-0.2184***	-0.6572***	-0.0762	-0.0536***	-0.0813***	-0.0838***
	-0.0129	-0.0404	-0.1893	-0.3396	-0.0158	-0.0191	-0.0253
Long Recall X Proxy	-0.4166***	-0.3948***	-0.8119***	-0.1619	-0.4210***	-0.0317	-0.0489**
	-0.0418	-0.0556	-0.2168	-0.2762	-0.0423	-0.0205	-0.0206
Mean Self-Report	0.9972	1.4068	4.4312	4.8469	0.9831	0.1102	0.1723
Mean Proxy	0.7762	0.8238	2.2133	1.6595	0.7381	0.0667	0.0762
Scaled Difference Self-Report	-0.0535	-0.1553	-0.1483	-0.0157	-0.0545	-0.7381	-0.4861
Scaled Difference Proxy	-0.5368	-0.4793	-0.3668	-0.0976	-0.5704	-0.4757	-0.6421
P-Val: Self=Proxy	0	0.0075	0.497	0.827	0	0.0656	0.2724
Observations	1167	1167	1167	1167	1167	1167	1167
R-squared	0.3969	0.3182	0.2104	0.1495	0.3916	0.0407	0.0404

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.

4.3 Heterogeneity by Gender and Age

While the initial evidence suggests that proxy respondents are more affected by the long recall window than respondents, making direct comparisons in estimated long recall loss between self and proxy reported labor measures is a challenge. There are likely to be major differences in the behavior underlying long recall loss when reporting about oneself versus someone else. Further, as acknowledged earlier, household heads are very different from other adult household members and are therefore likely to be engaged in different amounts and types of work. Still, to make progress understanding heterogeneity, we examine how long recall loss differs by gender and age within respondent type (self or proxy).

First, we examine heterogeneity by gender, focusing on the impacts of long recall on total activities and number of months worked in Table 4, with worked at all and hours worked reported in Appendix Table 2. For each outcome we estimate three specifications testing for different types of heterogeneity: (1) respondents (self-reports) only, considering heterogeneity by their own gender, (2) proxied individuals only, considering heterogeneity by the proxied individual's gender, and (3) proxied individuals only considering the respondent's gender. We do not observe any meaningful differences in long recall loss by gender either among self-reported measures in columns 1 and 4 or proxied measures in columns 3 and 6. Losses among proxied women appear to be larger in absolute and relative terms, especially for number of months worked in column 5, but the estimates lack precision to distinguish these differences from statistical noise. Our lack of differential effects by gender runs counter to those observed by Kilic et al. (2020).

Finally, we examine the patterns of effects by age group in Table 5 and Appendix Table 3, replicating the specifications and table structure from the gender analysis. While tests for differences between all four age groups do not always show statistical significance, it appears that older respondents have higher levels of long recall loss when reporting their own labor activities. Although column 3 does not show clear patterns, column 6 suggests that this pattern may be reversed when reporting about other household members, with older respondents exhibiting less long recall loss than younger respondents. Finally, younger household members appear to be more affected by long recall loss in both absolute and relative magnitude with significant differences among those under 25 or 25-34 when tested against people 50 and above.

Table 4: Long recall losses: Heterogeneity by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
		Total activities			Number of months worked	
Long Recall x Male	-0.2104***	-0.3320***	-0.3956***	-0.6544***	-0.3129	-0.8353***
	-0.0475	-0.1093	-0.0623	-0.22	-0.4582	-0.2445
Long Recall x Female	-0.2434***	-0.4144***	-0.3875***	-0.6658*	-0.9670***	-0.6761
	-0.0756	-0.063	-0.1181	-0.3692	-0.2348	-0.4636
Sample	Respondents	Proxy	Proxy	Respondents	Proxy	Proxy
Characteristics	Own	Own	Respondent	Own	Own	Respondent
Mean Male	1.4291	0.76	0.8483	4.4382	1.9459	2.3199
Mean Female	1.3441	0.8438	0.6875	4.4113	2.2969	1.6206
Scaled Difference Male	-0.1472	-0.4368	-0.4663	-0.1474	-0.1608	-0.36
Scaled Difference Female	-0.1811	-0.4911	-0.5637	-0.1509	-0.421	-0.4172
P-Val: Male=Female	0.7118	0.5049	0.9521	0.9788	0.1903	0.764
Observations	696	471	471	696	471	471
R-squared	0.0486	0.1152	0.1257	0.0221	0.06	0.0674

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.

Table 5: Long recall losses: Heterogeneity by Age Group

	(1)	(2)	(3)	(4)	(5)	(6)
		Total activities			Number of months worked	
Long Recall x Under 25	-0.1667	-0.4338***	-0.3043	-0.5327	-1.2248***	-1.3089*
	-0.1463	-0.0875	-0.19	-0.6064	-0.2942	-0.7136
Long Recall x 25 - 34	-0.0896	-0.4621***	-0.5341***	-0.5289	-0.9189**	-1.3946***
	-0.0805	-0.1004	-0.106	-0.3644	-0.4092	-0.3607
Long Recall x 35 - 49	-0.2470***	-0.3475***	-0.3665***	-1.0202***	-0.6037	-0.4466
	-0.0673	-0.1129	-0.091	-0.31	-0.4552	-0.3906
Long Recall x 50 plus	-0.3019***	-0.0656	-0.3034***	-0.3978	1.3433	-0.5994
	-0.0715	-0.204	-0.0986	-0.3582	-0.9405	-0.3975
Sample	Respondents	Proxy	Proxy	Respondents	Proxy	Proxy
Characteristics	Own	Own	Respondent	Own	Own	Respondent
Mean Under 25	1.3636	0.7701	0.8889	4.5313	2.0717	2.7037
Mean 25 - 34	1.3778	0.9483	1	4.4141	2.5572	2.6459
Mean 35 - 49	1.4228	0.8039	0.825	4.5461	2.2392	2.1693
Mean 50 plus	1.4202	0.7143	0.6452	4.3067	1.5746	1.7873
Scaled Difference Under 25	-0.1222	-0.5633	-0.3423	-0.1176	-0.5912	-0.4841
Scaled Difference 25 -34	-0.065	-0.4873	-0.5341	-0.1198	-0.3593	-0.5271
Scaled Difference 35 - 49	-0.1736	-0.4323	-0.4443	-0.2244	-0.2696	-0.2059
Scaled Difference 50 plus	-0.2126	-0.0919	-0.4703	-0.0924	0.8531	-0.3354
P-Val: U25=25 -34	0.6445	0.8309	0.2922	0.9957	0.5444	0.9153
P-Val: U25=35 - 49	0.6179	0.5346	0.7683	0.4743	0.2352	0.2891
P-Val: U25=50plus	0.4064	0.1031	0.9969	0.8481	0.0094	0.3861
P-Val: 25 - 34=35 - 49	0.1341	0.4474	0.2297	0.3048	0.6063	0.0765
P-Val: 25 - 34=50 plus	0.049	0.0851	0.1153	0.7976	0.0274	0.1429
P-Val: 35 - 49=50plus	0.5764	0.2298	0.6365	0.1893	0.0645	0.782
Observations	696	471	471	696	471	471
R-squared	0.0544	0.1206	0.1502	0.025	0.0741	0.0808

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.

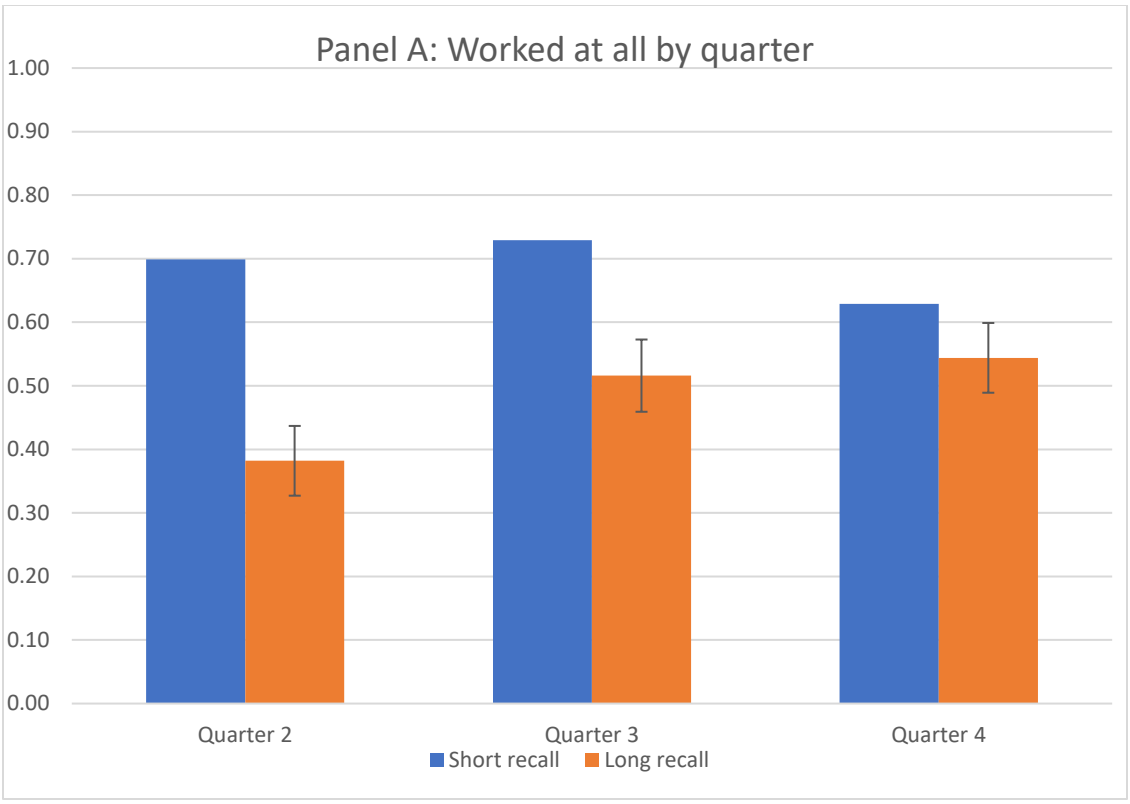
For both women and youth we note that, independent of heterogeneity in the effects of long recall, the fact that they are much more likely to be proxied than other household members is, itself, creating strong distortions towards the undercounting of their labor contributions.

4.4 Time pattern of results

To better understand the dynamics and mechanics of these results, we show results for the four main outcomes by quarter, again focusing on quarters 2, 3, and 4 in Figure 2. Consistent with our discussion in section 2, we omit quarter 1 as the long recall window did not sufficiently overlap with the first short run recall to afford comparable reference periods. Each panel shows the regression adjusted means by quarter, with a 95% confidence interval on the difference between the two recall groups.

Across all four outcomes, we see that long recall labor measures are furthest below the short run measures in the second quarter and the difference is highly significant, showing proportionate reductions of approximately 40%. However, reported differences in labor participation for the intensive measures, any work, and number of activities, grow smaller as the amount of time since the endline is reduced in quarters 3 and 4. The intensive measures, months and hours worked, show an even more extreme pattern whereby the long recall responses increase to a level significantly above the quarter 4 short recall estimates. We discuss this surprising pattern further in the next section, but the broad patterns of the results suggest that long recall appears to be leaving off considerable labor participation in the periods further from the endline, consistent with greater difficulty recalling and reporting activities that took place further in the past.

The patterns shown in these results are puzzling given that major activities over the previous 12 months should not be missed by the major activities that take place in 90-day segments within that range. One behavior leading to increasing labor measures over time in the long recall is if respondents have a tendency to “pull forward” the attributed months of a given activity from the months they report starting it, up to the present day at the time of the endline (in quarter 4). A second possibility would be if, when reporting typical days worked per month and typical hours worked per day, respondents instead give these responses with reference to heavier (and possibly more salient) months of activity. This behavior would be consistent with results seen elsewhere in the literature such as those by Arthi et al. (2018). If this is the case, we could similarly observe overestimation of labor measurements in the lower intensity months.



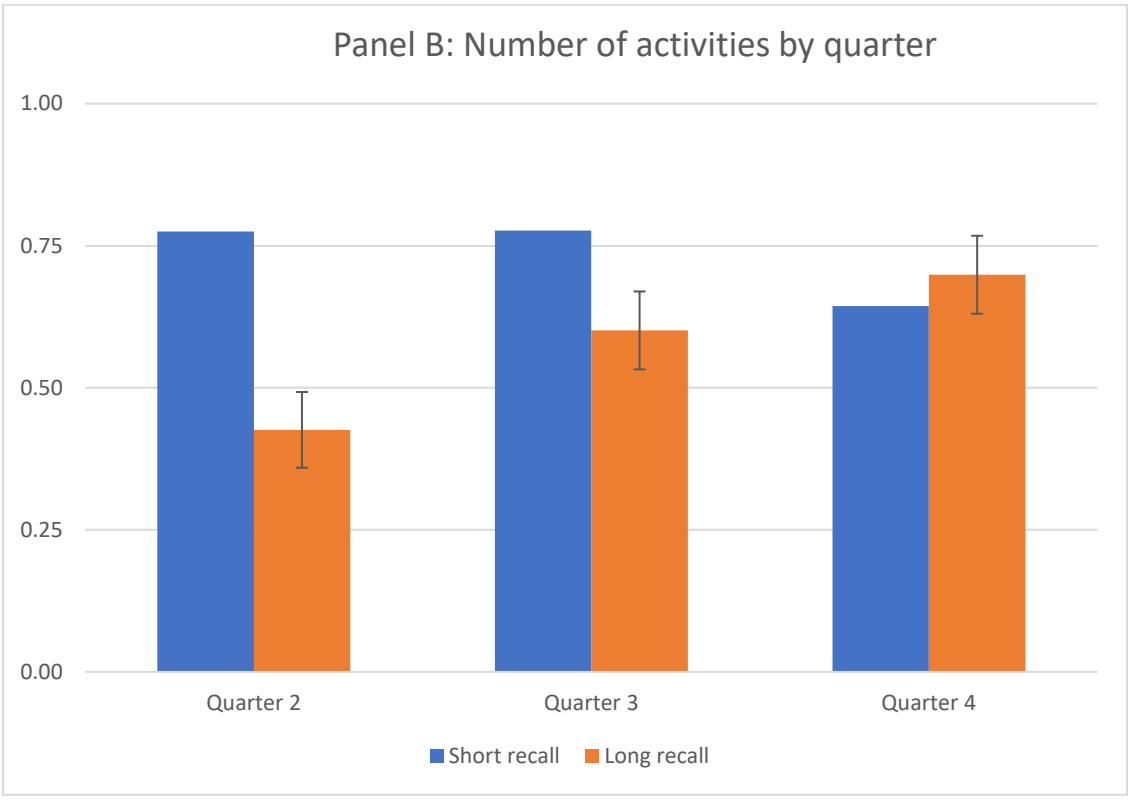
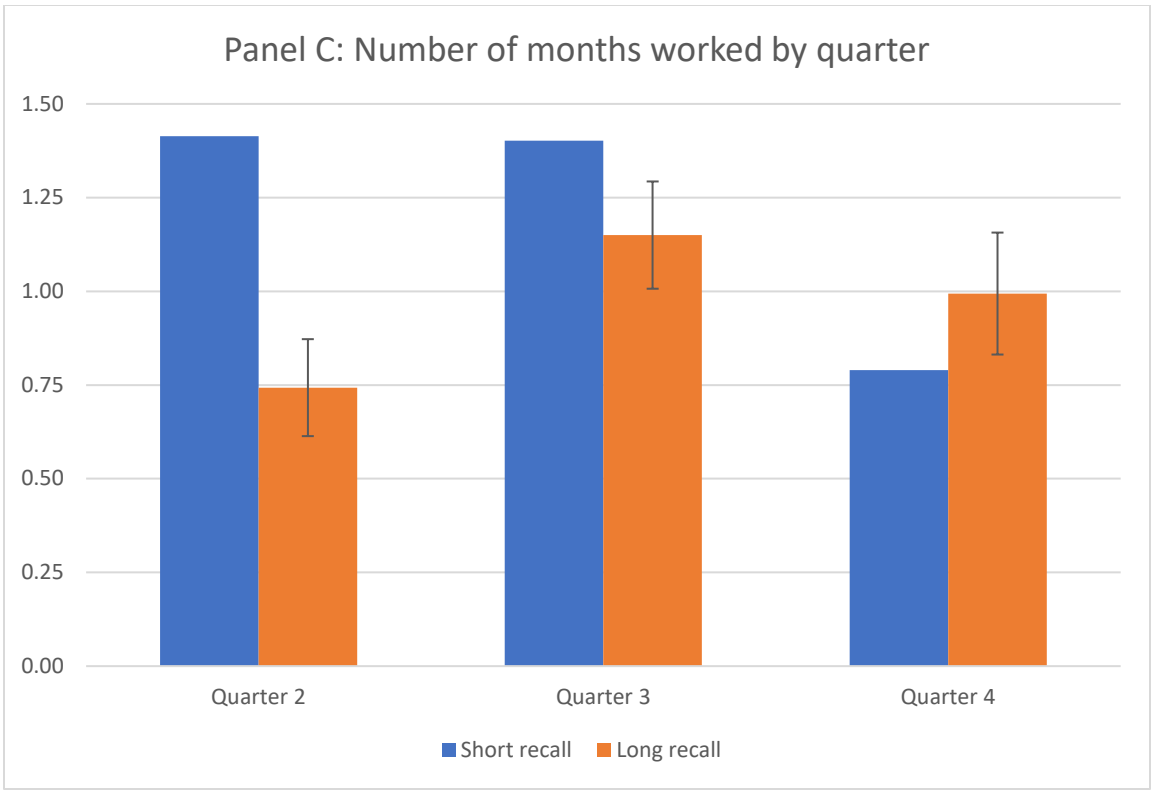


Figure 2: Labor supply and recall by quarter



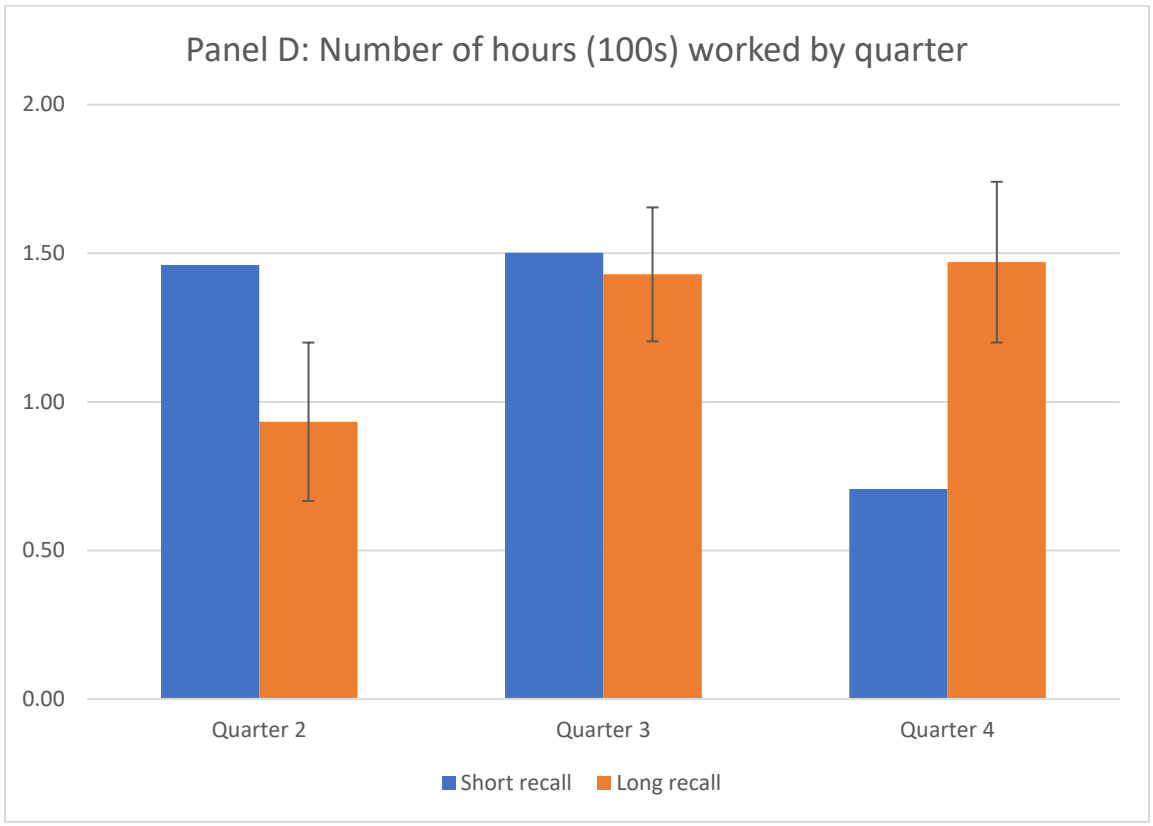


Figure 2: Labor supply and recall by quarter (cont'd)

5. Discussion

The results presented in this paper suggest that using an annual recall window to measure employment can lead to considerable losses relative to use of a shorter, quarterly, recall window. The extent of these losses are heavily influenced by whether an individual's labor contribution is being reported by themselves or being reported by someone else. Labor contributions of youth are more likely to be omitted when using longer recall windows and relying on proxy reports. The greater likelihood of women and youth to be proxied for in household surveys means that use of long recall periods may especially undercount their labor contributions. The time pattern of results are suggestive of more underreporting the further in time from the last survey round, coupled with possible overestimation of hours worked.

The majority of current research suggests that reported data is less accurate with longer recall windows, however even the shorter, three-month recall window that we use as our benchmark may be influenced by distortions from "truth". Three month recall may, itself, be missing meaningful activities that could be captured using even shorter interview intervals and recall periods. However, repeated surveys that aim to describe labor across a season or year must weight these potential gains against both survey costs and respondent fatigue. This work suggests that there are advantages to the three-month windows although future research could evaluate their performance against even shorter recall periods to further inform these tradeoffs. Additionally, recall windows of any duration could be affected by telescoping, leading to overestimation of reported measures. Adding participation across quarterly surveys could therefore disproportionately impact the short recall estimates although requiring individuals to report the specific months in which people worked likely reduced this vulnerability. Careful documentation of the months in which work occurred and the possible use of the previous survey round as a reference point may act to mitigate telescoping.

The results in this paper show very large drops in measurement of employment with long recall compared to a shorter recall period. This includes both measures for working at all, and for the number of jobs engaged in, central for characterization of rural income diversification. The extent to which we understand the time pattern of work also changes. This is not a principal goal of many multitopic surveys that often aim to estimate only the number of months worked and not which months. However, as a focus on the seasonality of work becomes more common, an

accompanying understanding of how to measure this seasonality are becoming increasingly important.

As policy-makers increasingly demand data driven insights, continuing to refine and improve our methods of data collection are only becoming more important. Understanding the seasonality and intensity of labor are central to rural development planning. But these themes cannot be separated from expanding our understanding of the effects of proxy reporting and recall windows on the data themselves. This effort is essential to reduce the risk that researchers end up inadvertently biasing their data through the very processes used to collect it.

References

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

Arthi V, Beegle K, De Weerdt J, Palacios-López A. 2018. Not your average job: measuring farm labor in Tanzania. *Journal of Development Economics.* 130:160–72

Asfaw, S., Scognamillo, A., Di Caprera, G., Sitko, N. and Ignaciuk, A., 2019. Heterogeneous impact of livelihood diversification on household welfare: Cross-country evidence from Sub-Saharan Africa. *World Development*, *117*, pp.278-295.

Backiny-Yetna P, Steele D,Djima IY. 2017. The impact of household food consumption data collection methods on poverty and inequality measures in Niger. *Food Policy* 72:7–19

Bardasi E, Beegle K, Dillon A, Serneels P. 2011. Do labor statistics depend on how and to whom the questions are asked? Results from a survey experiment in Tanzania. *World Bank Economic Review.* 25(3): 418–47

Beegle, Carletto, Himelein 2012a. Reliability of recall in agricultural data. *Journal of Development Economics*. 98: 34-41.

Beegle, de Weerdt, Friedman, Gibson 2012b. Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics.* 98: 3-18

Das J, Hammer J, Sánchez-Páramo C. 2012. The impact of recall periods on reported morbidity and health seeking behavior. *Journal of Development Economics.* 98(1): 76–88.

DeWeerdt J, Beegle K, Friedman J, Gibson J. 2016. The challenge of measuring hunger through survey. *Econ. Dev. Cult. Change* 64(4):727–58

De Weerdt, Gibson, and Beegle. 2020. What Can We Learn from Experimenting with Survey Methods? *Annual Review of Resource Economics*. 431-47

Desiere, Sam, Costa, Valentina, 2019. Employment Data in Household Surveys: Taking Stock, Looking Ahead. The World Bank.

Di Maio M, Fiala N. 2019. Be wary of those who ask: a randomized experiment on the size and determinants of the enumerator effect.*World Bank Econ. Rev.*

Dzanku, Fred Mawunyo, 2020. Poverty reduction and economic livelihood mobility in rural sub-saharan Africa. J. Int. Dev.

Djurfeldt, Agnes Andersson, 2013. African re-agrarianization? Accumulation or pro-poor agricultural growth? World Dev. 41, 217–231.

Ellis, Frank, 1998. Household strategies and rural livelihood diversification. J. Dev. Stud. 35 (1), 1–38.

Ellis, Frank, Freeman, H. Ade, 2004. Rural livelihoods and poverty reduction strategies in four African countries. J. Dev. Stud. 40 (4), 1–30.

Garlick, Orkin, Quinn. 2020. Call Me Maybe: Experimental Evidence on Frequency and Medium Effects in Microenterprise Surveys. *World Bank Economic Review.* 34(2): 418-443.

Haggblade, Steven, Hazell, Peter, Reardon, Thomas, 2010. The rural non-farm economy: Prospects for growth and poverty reduction. World Dev. 38 (10), 1429–1441.

Heath, Mansuri, Rijkers, Seitz, Sharma. 2021. Measuring Employment: Experimental Evidence from Urban Ghana. *World Bank Economic Review.* 35(3): 635-651.

Himanshu, Lanjouw, Peter, Murgai, Rinku, Stern, Nicholas, 2013. Non-Farm Diversification,Poverty, Economic Mobility and Income Inequality: A Case Study in Village India. The World Bank.

Imai, Katsushi S., Gaiha, Raghav, Thapa, Ganesh, 2015. Does non-farm sector employment reduce rural poverty and vulnerability? Evidence from Vietnam and India. J.Asian Econ. 36, 47–61.

Kilic, Van den Broeck, Koolwal, Moylan 2020. Are You Being Asked? Impacts of Respondent Selection on Measuring Employment. *World Bank Policy Research Working Paper* 9152.

Muyanga, Milu, Zephania Nyirenda, Yanjanani Lifeyo & William J. Burke. 2020.The Future of Smallholder Farming in Malawi. Working Paper No. 20/03. Lilongwe, Malawi: MwAPATA Institute (Accessed, December 19, 2021) https://www.mwapata.mw/_files/ugd/dd6c2f_f3cd0a352667458ea4e7ddd894db4ab3.pdf?index=true

Serneels, Beegle, Dillon. 2017 Do returns to education depend on how and whom you ask? *Economics of Education Review.* October Vol. 60. pp.5-19

Yeboah, Felix Kwame, Jayne, Thomas S., 2018. Africa's evolving employment trends. J. Dev. Stud. 54 (5), 803–832.

Appendix Table 1: Test of Losses from Order of Modules

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Three month recall				*Twelve month recall*			
	Worked at all	Number of months worked	Total activities	Hours worked (100s)	Worked at all	Number of months worked	Total activities	Hours worked (100s)
Asked first	0.0604***	0.0705**	0.3239**	-0.0394	0.2930***	0.3674***	1.7339***	2.1793***
	-0.0172	-0.0333	-0.1319	-0.2059	-0.0259	-0.0393	-0.1715	-0.2555
Mean in Reference Group	0.8391	1.0846	3.1566	3.5323	0.6932	0.864	2.7596	3.3672
Scaled Difference	0.072	0.065	0.1026	-0.0111	0.4226	0.4252	0.6283	0.6472
Observations	1167	1167	1167	1167	1167	1167	1167	1167
R-squared	0.2338	0.2437	0.239	0.1969	0.3275	0.2687	0.1953	0.1405

Notes: Columns 1 - 4 show the impact of asking the quarterly recall first on the quarterly recall measures. Columns 5 - 8 show the impact of asking the long recall questions first on the long recall measures. Other specification notes are as in the main tables.

Appendix Table 2: Long recall losses: Heterogeneity by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Worked at all			Hours worked (100s)		
Long Recall x Male	-0.0496***	-0.3412***	-0.4177***	-0.217	0.1968	-0.1455
	-0.0144	-0.0885	-0.046	-0.388	-0.5051	-0.3138
Long Recall x Female	-0.0647**	-0.4401***	-0.4086***	0.3591	-0.2735	-0.2433
	-0.0281	-0.0469	-0.105	-0.698	-0.3178	-0.5614
Sample	Respondents	Proxy	Proxy	Respondents	Proxy	Proxy
Characteristics	Own	Own	Respondent	Own	Own	Respondent
Mean Male	0.9962	0.72	0.7921	4.9142	1.37	1.7247
Mean Female	1	0.7937	0.6875	4.6578	1.7499	1.2967
Scaled Difference Male	-0.0498	-0.4739	-0.5274	-0.0442	0.1436	-0.0844
Scaled Difference Female	-0.0647	-0.5545	-0.5943	0.0771	-0.1563	-0.1876
P-Val: Male=Female	0.6322	0.3195	0.9368	0.4709	0.4185	0.8802
Observations	696	471	471	696	471	471
R-squared	0.0311	0.1991	0.2069	0.0058	0.0243	0.0248

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.

Appendix Table 3: Long recall losses: Heterogeneity by Age Group

	(1)	(2)	(3)	(4)	(5)	(6)
		Worked at all			Hours worked (100s)	
Long Recall x Under 25	-0.0357	-0.4454***	-0.3557**	-0.8953	-0.7601**	-0.7474
	-0.0353	-0.0668	-0.1664	-0.9922	-0.2952	-1.0236
Long Recall x 25 - 34	-0.0331*	-0.4700***	-0.4930***	0.2867	-0.154	-1.2546***
	-0.0189	-0.0733	-0.0727	-0.6686	-0.6315	-0.4332
Long Recall x 35 - 49	-0.0511**	-0.3565***	-0.3681***	-0.8718	0.2974	0.631
	-0.0205	-0.0861	-0.0723	-0.5494	-0.552	-0.5455
Long Recall x 50 plus	-0.0756***	-0.2534	-0.4039***	0.6599	1.7135	0.0243
	-0.0285	-0.1597	-0.0772	-0.6438	-1.4414	-0.4547
Sample	Respondents	Proxy	Proxy	Respondents	Proxy	Proxy
Characteristics	Own	Own	Respondent	Own	Own	Respondent
Mean Under 25	1	0.7241	0.8889	5.2029	1.4917	2.1664
Mean 25 - 34	1	0.8966	0.8983	4.9717	2.0077	2.1176
Mean 35 - 49	1	0.7255	0.7625	5.2672	1.6214	1.6176
Mean 50 plus	0.9916	0.7857	0.6613	4.2522	1.3979	1.2039
Scaled Difference Under 25	-0.0357	-0.6151	-0.4001	-0.1721	-0.5095	-0.345
Scaled Difference 25 -34	-0.0331	-0.5242	-0.5489	0.0577	-0.0767	-0.5925
Scaled Difference 35 - 49	-0.0511	-0.4914	-0.4828	-0.1655	0.1834	0.3901
Scaled Difference 50 plus	-0.0762	-0.3226	-0.6108	0.1552	1.2258	0.0202
P-Val: U25=25 -34	0.9486	0.8023	0.4505	0.3235	0.3832	0.6522
P-Val: U25=35 - 49	0.707	0.4054	0.9452	0.9835	0.0833	0.2279
P-Val: U25=50plus	0.3802	0.2767	0.7928	0.189	0.0931	0.4856
P-Val: 25 - 34=35 - 49	0.520	0.3151	0.2233	0.1811	0.5903	0.0076
P-Val: 25 - 34=50 plus	0.2154	0.2259	0.408	0.6878	0.2333	0.0485
P-Val: 35 - 49=50plus	0.4853	0.5738	0.735	0.0708	0.3589	0.3942
Observations	696	471	471	696	471	471
R-squared	0.0333	0.2012	0.235	0.0112	0.0346	0.0512

Notes: All estimates cover three quarters of data. Reported means are of short recall estimates. Scaled differences are the coefficients divided by the short recall mean. Standard errors are clustered at the household level.