Journal of Development Economics 130 (2018) 160–172

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Not your average job: Measuring farm labor in Tanzania

Vellore Arthi [a], Kathleen Beegle [b,*], Joachim De Weerdt [c], Amparo Palacios-López [b]

[a] University of Essex, UK
[b] World Bank, United States
[c] University of Antwerp and KU Leuven, Belgium

ARTICLE INFO

JEL codes:
C8
O12
Q12

Keywords:
Recall error
Measurement error
Farm labor
Agricultural productivity

ABSTRACT

Understanding the constraints to agricultural growth in Africa relies on the accurate measurement of smallholder labor. Yet, serious weaknesses in these statistics persist. The extent of bias in smallholder labor data is examined by conducting a randomized survey experiment among farming households in rural Tanzania. Agricultural labor estimates obtained through weekly surveys are compared with the results of reporting in a single end-of-season recall survey. The findings show strong evidence of recall bias: people in traditional recall-style modules reported working up to four times as many hours per person-plot as those reporting labor on a weekly basis. Recall bias manifests both in the intensive and extensive margins of labor reporting: while hours are over-reported in recall, the number of people and plots active in agricultural work are under-reported. The evidence suggests that this recall bias is driven not only by failures in memory, but also by the mental burdens of reporting on highly variable agricultural work patterns to provide a typical estimate. All things equal, studies suffering from this bias would understate agricultural labor productivity.

1. Introduction

Of the 1.4 billion people living in extreme poverty, the majority reside in rural areas and rely on agriculture as a source of income and livelihood (Olinto et al., 2013). In Sub-Saharan Africa, nearly 75 percent of the extreme poor reside in rural areas, and over 90 percent participate in agriculture. Smallholder agriculture is the predominant form of farm organization, with 33 million small farms holding less than two hectares and representing 80 percent of all farms in Africa (FAO, 2009). On these farms, agricultural practices are typically labor intensive, and the majority of the labor is provided by household members.

Accordingly, the labor of household members in agriculture is a key asset for poor households, and its accurate measurement is essential to the development of sound policy. Despite the importance of the agricultural sector in reducing poverty and food insecurity (Chen and Ravallion, 2007; Irz et al., 2001; Ligon and Sadoulet, 2007), serious weaknesses in agricultural statistics persist.[1] In this study, we examine one aspect of this issue: measures of family farm labor. Specifically, we test for bias related to the length of the recall period over which labor must be reported.

To assess the degree of recall bias in household farm labor, we conducted a survey experiment in Mara Region, Tanzania, over the long rainy season, January–June 2014. Smallholder farming households were randomly assigned to one of four survey designs, varying the mode (face-to-face versus phone) and frequency of interview, and, thereby, the recall period. Household labor information collected in weekly visits—our resource-intensive gold standard—is then compared with data reported after the harvest. After establishing the magnitude of recall bias, we investigate the mechanisms by which it arose.

We find recall bias in the reporting of family farm labor; however, because of competing forms of recall bias in the reporting of hours of labor, the number of plots, and the number of farming-active household members, the degree of distortion in reporting depends on the level of data aggregation. Labor data collected on a weekly basis, whether in person or by phone, are similar, albeit sometimes moderately statistically different. There are, however, striking and economically meaningful differences between the weekly and recall data. Respondents in recall-style modules report working up to nearly four times as many hours per person per plot, compared with respondents reporting labor on a weekly basis. Meanwhile, recall-surveyed households under-report both the number of household members and plots active in farm cultivation. Evidence suggests that these sources of recall bias are driven not only by

* Corresponding author.
E-mail address: kbeegle@worldbank.org (K. Beegle).

[1] See ABCDQ (Agricultural Bulletin Board on Data Collection, Dissemination, and Quality of Statistics) (database), Statistical Division, Food and Agriculture Organization of the United Nations, Rome, http://faostat.fao.org/abcdq/.

https://doi.org/10.1016/j.jdeveco.2017.10.005
Received 11 October 2016; Received in revised form 8 October 2017; Accepted 16 October 2017
Available online 23 October 2017
0304-3878/

failures in memory where farm inputs are non-salient, but also by the mental burdens of computing data on a typical situation if agricultural work patterns are highly variable during the season.

Our results have important implications for development policy and fill key gaps in the literature concerning survey methods and the quality of agricultural labor data. Ours is one of the few studies to test the accuracy of agricultural labor data in developing-country settings. While labor data have been an essential ingredient in a broad range of important studies on smallholder agriculture in developing countries, scant attention has been paid thus far to the quality and robustness of the underlying data on family farm labor. Evidence that agricultural labor inputs may be substantially overestimated calls into question the reliability of the traditional end-of-season labor estimates commonly collected in household surveys measuring such labor.

These findings also contribute to academic and policy debates concerning the agricultural productivity gap and the degree to which rural labor may be misallocated in developing economies. Several studies have been engaged in this debate. Two in particular, Gollin et al. (2014) and McCullough (2016), question the accuracy of current labor measures and reconsider the agricultural productivity gap after adjusting for labor data quality. By conducting comparisons at the per-hour level (McCullough, 2016) and by adjusting for sectoral differences in hours worked as well as for levels of human capital (Gollin et al., 2014), both studies find that the difference in the productivity between the agricultural and non-farming sectors is narrower than usually thought. Our study suggests that surveying irregular labor through recall may result in an upward bias in the reported hours of farm labor, which would further help explain this productivity gap.

Although our results call into question the accuracy of current farm labor data, they also suggest specific ways to improve the accuracy of labor measurement. For instance, the consistency of labor reporting across face-to-face and phone surveys suggests that season-long phone surveys are one option for reducing error in the measurement of rural agricultural labor.

The rest of the paper proceeds as follows. In Section 2, we offer background on labor measurement. In Section 3, we provide an overview of our empirical approach, including details on the survey experiment. In Section 4, we present the results and outline the mechanisms by which bias manifests in recall data through both the extensive and intensive margins of labor reporting. Section 5 concludes.

2. Measuring labor

2.1. Current practice

The wealth of evidence on the quality and reliability of labor statistics in household surveys comes largely from the United States (for a thorough review, see Bound et al., 2001). In developing and agriculturally-driven countries, for contrast, little is known about the extent to which the design of surveys influences labor statistics. Clearly, it is difficult to extrapolate from studies conducted in the United States to the African context. Moreover, the existing literature on data quality and survey methods in low-income settings rarely pertains to farm labor (see Bardasi et al., 2011). It has been noted that International Labour Organization recommendations for measuring labor are likely to be inadequate in settings such as rural Tanzania, where the majority of labor is found in the informal, self-employed, and farm sectors (World Bank, 2014).

Our review of over 35 recent household surveys that collect labor data in Africa shows that, in practice, the capture of labor market statistics in household surveys varies widely. The recall period, the sequencing of questions, the use of screening questions, the seasonal timing, the granularity of reporting requested, the unit over which labor is reported, and the choice of respondent can vary across surveys both within and across countries. Differences in household survey design have been shown to have substantial implications for statistics and analysis of welfare, poverty, and hunger (Backiny-Yetna et al., 2014; Beegle et al., 2012b, 2016; De Weerdt et al., 2016), as well as labor measurement (Bardasi et al., 2011) and a range of other socioeconomic conditions.

National integrated or multi-topic household surveys in Africa generally collect data on agricultural labor in two ways.[2] In one approach, general labor information, including agricultural labor, is collected in a labor module. In another, specific agricultural labor data are collected in a dedicated agriculture module, such as in the Living Standards Measurement Study–Integrated Surveys on Agriculture (LSMS-ISA). In the former case, information on labor involving each household member above some specified age is collected in reference to the last seven days or, perhaps, the last 12 months (Anderson Schaffner, 2000). The person's labor input is not differentiated by plot, by crop, or by farm activity (such as weeding, harvesting, and so on). Instead, in the agricultural module outlined by Reardon and Glewwe (2000), the total days of labor at the household level over the last completed season are collected for each plot and by specific farming activity. An expanded agricultural module would have the same questions for each household member (as in the LSMS-ISA).[3] A common feature in these surveys is that labor information is collected from a single interview.

Though they are considered an improvement over surveys with more general labor force questions, surveys like the expanded LSMS-ISA agricultural module have several potential drawbacks. First, it is time-consuming to collect this very detailed information. Second, the burden on respondents is substantial: respondents are asked to provide information that they may never have considered (for instance, about labor by activity for each plot). Third, there is potential for problems in recall and memory. In our study, we show that these last two points in particular may contribute to inaccuracy in farm labor reporting.

2.2. What complicates the measurement of smallholder farm labor?

2.2.1. Features of smallholder farming

The estimation of labor inputs on smallholder farms is complex and vulnerable to misreporting.[4] Smallholder farms typically employ mostly family labor, and so there is no wage income on which to anchor recall. Written records are rarely kept, and the respondent must rely on recall strategies to report on past events. To arrive at the total amount of labor allocated by a household to farming, the household must accurately report the plots under cultivation, the specific household members who worked on each plot, the activities performed, and the timing and duration of these activities. Farming is a seasonal activity, and work patterns are irregular during the season. Reporting on the typical or average amount of time spent farming requires, after the completion of the season, remembering distant events and performing complicated mental calculations. Alternatively, reporting hours worked in the last seven days at any single point during the agricultural season will not necessarily be indicative of total labor during the season, if labor inputs vary greatly during the season—particularly if respondents report on what “typically” happens in a given week, rather than what actually happened in the preceding week. Accordingly, farm labor measures can

[2] Apart from multi-topic household surveys, smallholder information can be collected through specialized farm surveys. These often entail visiting the household at multiple times, particularly those surveys utilizing resident enumerators (for example, agricultural extension agents or other ministry of agriculture staff). However, these surveys typically do not collect details on household farm labor.

[3] The LSMS-ISA program has been conducted in Burkina Faso, Ethiopia, Malawi, Mali, Niger, Nigeria, Tanzania, and Uganda. See LSMS (Living Standards Measurement Study) (database), World Bank, Washington, DC, http://www.worldbank.org/lsms.

[4] Measurement problems are not restricted to labor. For instance, intercropping, continuous planting, extended harvest periods, and multiple plots of small sizes and irregular shapes can make reporting on most inputs and outputs difficult. Although several strategies are proposed in the literature to account for mixed-stand crops, no method has yet gained wide acceptance (Fermont and Benson, 2011). The introduction of Global Positioning System devices has improved the measurement of landholdings, but the methods for collecting production and input data are not much different now than in the last several decades (Deininger et al., 2011).

suffer from recall issues at both at the extensive (e.g., plots and individuals active in farming) and intensive (e.g., number of weeks, days, or hours worked, conditional on working) margins.

2.2.2. Insights from cognitive psychology

In addition to issues arising from the complexity of smallholder farming patterns, the design of the survey instrument itself may also influence the quality of data on family farm labor. Considering common survey practices and the features of smallholder farm labor, alongside insights from the social and cognitive psychology literature, there is a particular need for caution in interpreting farm labor data taken from household surveys.

Perhaps the most important aspect in our context is the implications of the recall period. These recall effects can operate firstly through faults in memory. Forgetting an event is more likely as time passes. Alternatively, telescoping, by which a respondent remembers a distant event as if it occurred more recently, can result in memory-driven distortions, particularly in longer recall periods (Sudman and Bradburn, 1973). An example is a respondent who last worked on the farm 35 days ago, but who nevertheless reports that he worked on the farm within the past 30 days. Beegle et al. (2012a) find little evidence that longer recall periods lead to less reliable reporting of hired farm labor in Kenya, Malawi, and Rwanda, but less is known about the reliability of reporting on own-household labor, for which written records are less prevalent.

The length of the period of recall in survey responses may be important beyond the implications of memory processes: it can affect how a respondent interprets the questions. In asking about episodes of anger, Brown et al. (2007) found that if the recall period is one day, the respondent assumes that minor irritations should be counted. Extending the recall period to one year leads the respondent to believe that only serious incidents of anger should be reported. The shift in inferred pragmatic meaning makes it difficult to disentangle the effects of question interpretation and the effects of forgetting. In our context, asking about labor over a very extended period might lead respondents to omit reporting the modest time they spend on small plots or incidental crops, which might otherwise be included when respondents are asked about the last week. Das et al. (2012) find a similar pattern in the self-reporting of past health, whereby smaller illness events are ignored or forgotten as the recall period increases. They also find heterogeneity in these effects by income, driven by the normalization by the poor of what would otherwise—that is, for richer people—be salient illness events worthy of medical treatment. In our context, a farmer may interpret the question differently if asked to report on labor in the last week, as compared with someone who is asked to report on several months' worth of labor at the end of the season. Our results suggest that even seemingly straightforward questions, such as how many plots the farmer has cultivated, or who has worked on them, are affected by the recall period.

Beyond the length of the recall period, there are aspects of the cognitive and communicative processes that affect survey responses. Menon (1993) shows that for infrequent and salient events, respondents are likely to recall and count individual instances of these events because they are stored episodically and remain in memory for a longer time. In the absence of episodic event information that is easily retrieved, respondents will rely on other strategies. For less salient but very regular events, such as “I visit my grandmother every Saturday,” respondents are not likely to use the recall-and-count strategy, relying instead on the information they have stored about the event's periodicity. Such rate-based estimations may be adjusted by memories of nonoccurrence (“except when I'm on holiday”) or more frequent occurrence (“also on her birthday if that doesn't fall on a Saturday”). Menon (1993) notes that counting the occurrence of events that are neither salient nor regular requires much more cognitive effort on the part of the respondent. Thus, where work is neither salient nor regular, as may be the case for the labor of smallholder farmers over an agricultural season, respondents are unable to use rate-based or recall-and-count strategies and, so, are likely to yield erroneous reports of labor.

In the absence of episodic or rate-based information, respondents may revert to their general assumptions about the state of the world in their search for answers to survey questions. These assumptions then form a benchmark that is used to infer previous behavior. Indeed, the spuriously high recall-surveyed labor we find in our study can stem from this sort of inference. Schwarz and Oyserman (2001) cite evidence that retrospective estimates of income and of tobacco, marijuana, and alcohol consumption are unduly influenced by people's income and consumption habits at the time of the interview. Thus, they infer their previous behavior based on their current or recent behavior. Similarly, de Nicola and Giné (2014) show that survey responses on income from small-scale boat owners in coastal India rely more on inference, and less on true recollection, as the recall period increases. The authors show that, while this bias has little influence on the mean (because, in their case, fishermen base their inferences on average earnings), it does lead to an underestimation of income variability as the recall period increases. The information and assumptions held by respondents are also important if people report on the behavior of others, a common practice in the collection of labor data in household surveys (Bardasi et al., 2011; de Nicola and Giné, 2014).

Respondents may also be suggestible and base their inferences on what they believe *should have* occurred. For example, Ross and Conway (1986) allowed students to participate in a skills-training program that did not, in fact, influence their skills. After participating in the study, the students quantified their pre-training skills at a lower level than the level at which they had originally assessed their skills prior to receiving the skills training. The authors argue that the students reconstructed their past, guided by their subjective theories over what the skills training ought to have done. If African farmers hold implicit theories about the link between, say, labor inputs and production, then the report on the one may influence the report on the other. For example, in an end-of-season recall survey, labor may be retrospectively overstated during good harvests and understated during bad harvests. Thus, we might expect features of smallholder farming to exacerbate reporting issues generated by long recall periods—especially in a setting, like ours, where farm labor is irregular and non-salient, and so, where the cognitive burdens of reporting over long recall periods are high.

3. Experimental design and context

The goal of this study is to examine biases of the sort described above, in agricultural labor data collected through household surveys. We focus on potential biases introduced by the length of the recall period and the frequency of reporting. To do this, we conducted a large randomized survey experiment among smallholder farming households in rural Tanzania, through which we compare agricultural labor information collected in weekly surveys (our benchmark for the true labor estimates) with that collected in a single end-of-season survey. Here, we focus primarily on examining plot-person labor reporting. However, because understanding farm productivity at the lowest level entails studying inputs and yields on plots, and so may require analysis of aggregated measures, we also briefly touch on the reporting of aggregate household measures of farm labor.

3.1. Experimental design

We conducted a survey experiment among 854 farming households in 18 enumeration areas in the Mara Region of rural northern Tanzania. Labor input was measured for the 2014 *masika* (the main, long-rainy season), running roughly from January to June 2014. Households were randomly assigned to one of four survey designs within each of the 18 enumeration areas. The four survey arms differ in the manner and frequency with which they were contacted.[5]

[5] The data were collected using computer-assisted personal interviewing through the *surveybe* software program.

Two of the survey designs entailed weekly interviews throughout the entire *masika* season either in person or by phone, with face-to-face interviews at the start of the season and after the end of the agricultural season in July–September 2014.[6] The other two survey designs entailed a recall survey fielded after the end of the season. The four alternative survey designs are as follows:

- ***Weekly visit (benchmark):*** *Weekly face-to-face surveys for the duration of the masika*

  For weekly visit households, a baseline survey was conducted in January 2014, followed by weekly face-to-face surveys conducted by enumerators through the end of June 2014 and an endline survey (July–September 2014) to collect farm production information. For each plot, household members who had worked on the plot during the previous week were identified, and the hours for each day they worked on the plot during the previous week were reported.[7]
- ***Weekly phone:*** *Weekly phone surveys for the duration of the masika*[8]

  For weekly phone households, a face-to-face baseline survey was conducted in January 2014 (during which households were provided with mobile phones to respond to subsequent surveys), followed by weekly phone surveys through the end of June 2014 and a face-to-face endline survey in July–September 2014 to collect farm production information. For each plot, household members who had worked on the plot during the previous week were identified, and the hours for each day they worked on the plot during the previous week were reported.
- ***Recall NPS:*** *Face-to-face survey at the end of the masika, standard NPS module*

  For recall NPS households, a face-to-face endline survey was conducted after the harvest (July–September 2014), during which both labor and farm production information was collected. The agricultural labor module was identical to the respective module in the Tanzania NPS, waves 3 (2012/13) and 4 (2014/15). For each plot, the household members who worked the plot at any point during the season were identified, and the following information was reported: (a) total days spent on the plot over the season in each of four activities (land preparation and planting; weeding; ridging, fertilizer application, and other non-harvest activities; and harvesting) and (b) typical hours per day worked in each of these four activities.
- ***Recall alternative (ALT):*** *Face-to-face survey at the end of the masika, alternate survey module*

  For recall ALT households, a face-to-face endline survey was conducted after the harvest (July–September 2014), during which both labor and farm production information was collected. For each plot, the household members who had worked on that plot at any point during the season were identified, and the following information was reported: (a) total weeks worked on the plot over the season (irrespective of activity), (b) approximate number of days per week worked, and (c) approximate number of hours worked per day.

Throughout this paper, we report the magnitude of bias through comparisons with the weekly visit design.[9] This is based on the premise that the data reported in the weekly visit design are likely to be the closest to actual labor activities. We assume that the short one-week period and the specificity of the questions on farm labor reduce the influence of forgetting. Anchoring the reporting to the previous interview reduced the possibility of telescoping.

This choice of benchmark is validated by additional evidence on smallholder farming in East Africa. For instance, in addition to the quantitative surveys, in August 2016, we held focus group discussions in 5 communities from our original sample. In each community, there was a separate focus group for men and for women, consisting of 5 adults each. The semi-structured discussions delved into details about household structure and labor on and off the farm. The focus group discussions were purposively fielded in light of the large gaps found in preliminary analysis of the hours reported in recall and weekly surveys. The exercise was a means of independently confirming that the weekly surveys were indeed a reasonable benchmark of actual work.[10] These qualitative findings supported the use of the weekly data as a benchmark.

Before presenting the results of the survey experiment, there are several identification concerns with the study design that are worth noting. First, households were randomized within villages to account for micro agro-ecological patterns affecting household labor (which we may not capture through our other data sources). This raises the possibility of intra-cluster contamination, whereby one person's response is influenced by another's design status. We opted for within-village randomization because we believed that such contamination was unlikely because the villages in question were relatively large and diffuse.

Second, the weekly visits themselves could have influenced the labor decisions made by households (in a manner akin to Hawthorne effects). We cannot rule this out, but the evidence suggests that Hawthorne effects are unlikely to drive our results on recall bias. For instance, we do not find evidence of a significant seasonal trend in hours worked, except for an increase towards the harvest period. There was also little difference between the face-to-face and phone interviews, whereas one might expect Hawthorne effects to be stronger in in-person visits. Likewise, we find no systematic evidence of respondent fatigue in the weekly-surveyed households: the time it took to interview respondents is consistent with the intensity of work over the season, i.e., it is relatively constant across the season, with a rise during harvest.

Third, self-reporting rates were similar across survey designs. In the weekly visit group, interviewers were instructed, where possible, to collect information directly from respondents in order to avoid proxy reporting. Meanwhile, in the weekly phone interviews, one household member typically reported on his/herself as well as on other household members, although the possibility exists that people may have self-reported by turn. In both recall survey designs, and consistent with current common practice, interviewers were instructed to ask the most knowledgeable person in the household to report on family farm labor. Despite differences in the instructions given to enumerators and in the feasibility of self-reporting by survey type, the degree of self-reporting achieved was similar across the four survey designs. The response rates among self-reporting respondents were as follows: weekly visit (35 percent), weekly phone (33 percent), recall NPS (27 percent), and recall ALT (28 percent).

Finally, attrition was minimal. Households that were surveyed weekly and that dropped out within the first five weeks following the baseline

[6] All weekly visit households received a mobile phone, but recall households did not. Mobile phone ownership is widespread, at 72 percent of households in our sample. Thus, this element is unlikely to influence the results.

[7] In addition, after the hours per person per day over the previous week were reported, the range of activities performed during that time was recorded (land preparation and planting; weeding; ridging, fertilizer application, and other non-harvest activities; harvesting), but the number of hours were not specified for each activity.

[8] The weekly phone interview design draws on lessons summarized by Dillon (2012), who uses a phone survey to collect information on purchased input applications among cotton farmers in Tanzania. Similar recent work has used phone surveys to collect high-frequency data on economic activity. See Garlick et al. (2015) for a review of the literature on phone-based strategies for collecting household and enterprise data.

[9] In some parts of the analysis, however, we collapse the two weekly and two recall arms of the study for simplicity in comparisons.

[10] Unfortunately, we were unable to identify other sources of data on labor hours in similar farming systems collected intensively over an agricultural season to assess whether the weekly reporting is a valid benchmark. Queries to several agricultural economists who study small-holder farming systems in the region do not yield concrete estimates of labor inputs, although informal feedback was that our weekly reporting estimates seemed more plausible than the recall alternative.

Table 1
Sample characteristics.

	Weekly Visit	Weekly Phone	Recall NPS	Recall ALT
Individuals (*N* = 5375)				
Age	20.98 (20.12)	22.47* (20.47)	22.34* (20.70)	21.60 (19.71)
Proportion aged 10 years and over	0.63	0.67**	0.63	0.63
Proportion male	0.49	0.48	0.49	0.51
Proportion in school	0.28	0.32**	0.30	0.30
Proportion living with spouse	0.27	0.31*	0.28	0.27
Proportion literate	0.58	0.61	0.56	0.56
Proportion father deceased	0.28	0.26	0.29	0.28
Proportion mother deceased	0.16	0.17	0.17	0.16
Proportion visit health care provider past 4 weeks	0.16	0.15	0.15	0.14
Households (*N* = 854)				
Household size	6.4 (3.1)	6.5 (3.3)	6.3 (2.9)	6.2 (2.4)
Rooms in dwelling	2.9 (1.2)	3.1 (1.3)	2.9 (1.1)	3.0 (1.2)
Minutes to water source	58.5 (48.3)	55.0 (43.4)	54.8 (45.7)	53.5 (41.5)
Proportion with good walls	0.47	0.48	0.40	0.44
Proportion with good roof	0.74	0.78	0.76	0.78
Proportion with good floor	0.22	0.32**	0.24	0.31**
Number of households	212	212	212	218

Note: Table uses endline data. Mean values which are significantly different from the mean for the Weekly Visit group are denoted as follows: ***p < 0.01, **p < 0.05, *p < 0.1.

interview were replaced at random from the list of unassigned households. In the weekly visit group, 17 (7 percent) households surveyed in the baseline later dropped out of the study; these were replaced by 14 households, for a total of 212 weekly visit households reporting data for the main season. In the weekly phone group, 14 (6.2 percent) households dropped out, and 12 were added as replacements, for a total of 212 households reporting agricultural labor throughout the season. Replacements were made in this manner up to the sixth week of the weekly interviews. None of the recall-surveyed households declined to participate.

Table 1 presents descriptive statistics on household characteristics across the four survey designs, drawing on the endline (i.e., post-harvest) survey. For these set of traits, households were well balanced across the different survey designs. Jointly, these traits are not significantly different across designs. Of note, the household roster for the weekly visit and weekly phone households was collected slightly differently than for the recall households. For the former two groups, the roster was started at baseline and updated each week (identifying members who had left and new members since the previous week), and then again during the endline (a few weeks after the last weekly interview). We find no significant differences in the roster of household members at the endline between the weekly and recall households, whether in the total number of members or in the demographic profile of members.

3.2. Farming practices in Mara

Although its location on the edge of Lake Victoria enables a small fishing industry, Mara Region is primarily agricultural. The bulk of farming activity takes place over the main long rainy season (the *masika*), which runs roughly from January to June. The two main crops cultivated in the villages in our study are maize and cassava. Maize has a fixed seasonal cycle of land preparation, planting, weeding, and harvesting, a cycle which is governed by the onset of the rains.[11] By contrast, cassava has no specific cultivation cycle and is grown throughout the year. Cassava harvesting also occurs throughout the year, depending on household food needs, rather than at one specific point in time. Households frequently diversify cultivation, intercropping the two staples with beans, sweet potatoes, and sorghum.

Before comparing labor reporting by survey design, we use the benchmark weekly visit data to provide some context. Households had an average of 6.4 members with one-third of them children under 10. The average household cultivated 4.6 plots of about 1 acre each. These plots tended not to be located adjacent to the household's dwelling, nor were they typically adjacent to each other. On average, households reported their plots as being located a 26-min walk from the primary residence.[12]

Most people aged 10 or above were engaged in household farm labor. Table 2 provides an overview of the activities of these household members in our sample according to the weekly visit data. Consistent with the agricultural character of the region, the most common activity was work on a household farm; 88 percent of people spent at least one day in this activity over the season. Paid work, whether agricultural or otherwise, was rare: only 16 percent of people engaged in any paid agricultural work for others, and 11 percent performed paid nonagricultural work. A large share of people spent at least some time collecting firewood and water. About a quarter spent at least one day in school, and slightly less than half were sick for at least one day over the season.

Table 2, column 2 shows the average number of days spent in a given activity, as reported through the weekly visits, conditional on the performance of any reported labor activity that week. While important, family farm labor was perhaps less frequent than might be expected: people spent an average of 1.88 days a week working on their household farms, conditional on the reporting of any work that week. We show, however, that this does not necessarily imply a regular weekly work pattern. There was considerable irregularity and cyclicality in agricultural work. As suggested in a number of studies of farm labor in Sub-Saharan Africa (see the discussion in Arthi and Fenske, 2016), we find that the agricultural workday typically lasted four or five hours. This is much shorter than the hours spent in nonagricultural and market activities (such as paid non-agricultural work, non-agricultural household business, fishing, livestock keeping, and schooling), conditional on the performance of such work.

The largest portion of each workday was devoted to household agriculture. Fig. 1 gives an overview of the hours per day across activities as reported in the weekly visits. Fig. 1A averages across all people ages 10 or above for all days, and Fig. 1B excludes weekends and days the person was ill. Roughly a third of the total of 3.6–4.2 working hours,

[11] Our experiment was initiated at the beginning of the maize cycle, in January 2014, and followed respondents to the completion of the harvest in August–September.

[12] The time to commute to and from plots is not included in the working times reported in this study. Households were explicitly instructed to exclude commuting time in reporting the time worked in farming activities.

Table 2
Overview of Activities during the agricultural season.

Activity	Share of Individuals reporting the activity at least once over season	Average days per week in activity, conditional on reporting the activity at least once over the season	Hours per day in activity, conditional on activity that day
Household farm	0.88	1.88	4.49
Paid agricultural	0.16	0.34	4.65
Free agricultural, other hh	0.21	0.28	4.38
Fishing	0.10	1.24	6.38
Livestock	0.27	1.08	5.08
Paid non-agricultural	0.11	1.00	8.38
Non-agricultural business	0.31	1.43	7.59
Collecting firewood	0.56	0.49	2.01
Collecting water	0.73	2.72	1.23
Schooling	0.27	2.76	7.86
Sick	0.49	N/A	N/A

Note: The table is based on Weekly Visit data, and is restricted to individuals aged 10 years and over.

respectively, were devoted to agricultural activities. These data obscure important distributional differences, to which we return in later sections. Finally, Fig. 1C shows the allocation of time on days when at least some household farm activity was reported. On average, 5.8 hours were spent across all activities, of which 78 percent was spent on household farming. The remainder of the time was made up largely of collecting water, tending to livestock, and attending school.

4. Results

4.1. Main results

4.1.1. Intensive margin of misreporting: farm labor hours, conditional on any farm labor

To examine the implications of survey design on the reporting of household farm labor, we begin by looking at the intensive margin of farm labor reporting: the hours spent in farm labor, conditional on spending any time in farming. Specifically, we calculate the total number of hours spent in farm labor over the entire season, for every household plot-person combination where either the person or the plot in question was active in farming at any point during the agricultural season. Put another way, this measures the total number of hours of farm labor performed over the entire season by each household member active in agriculture on each household plot under cultivation. This calculation of total season hours (henceforth, person-plot hours[13]) is thus based on the most granular measure of labor inputs available in the survey.[14]

Table 3 reports mean total season hours at the person-plot level. Hours per day were exaggerated by roughly 7 percent in the recall surveys. The mean hours per day worked are consistent with reporting in the focus group discussions, where work was concentrated in morning hours (between 6 and 10am or 7–11am) and sometimes late afternoon (4–6pm).

Total weeks in recall are higher than in weekly visits by 128 percent, and total days are higher by 179–223 percent. The cumulative effect of the exaggerated days and weeks in the recall modules results in a striking recall bias-driven gap in the time spent by people working on a given plot in recall versus weekly surveys. For the weekly visits, the person-plot average of total season hours was 39.5; this number jumped to 121.3 and 146.3 in recall NPS and recall ALT, respectively. Total hours worked per person-plot are 3.0 and 3.7 times higher in the recall surveys than in our preferred benchmark, the weekly visit estimates.[15] There is considerable recall bias in season-wide person-plot hours, driven primarily by error in the least granular time unit reported (days in the case of the recall NPS, and weeks in the case of the recall ALT).

We will show in Section 4.2 that this pattern—in which hours worked per working day were reported more accurately than days and weeks worked—is consistent with the fact that hours worked per day were more regular and less variable, relative to weeks worked or days per week worked.

4.1.2. Extensive margin of misreporting: people and plots engaged in agriculture

Clearly, there is evidence that the total hours worked per person-plot is over-reported in recall modules relative to weekly data. Is there similar evidence of misreporting on the number of people and plots reported as active in agriculture? If so, then aggregating hours per person-plot over people, over plots, or over both could introduce further biases.

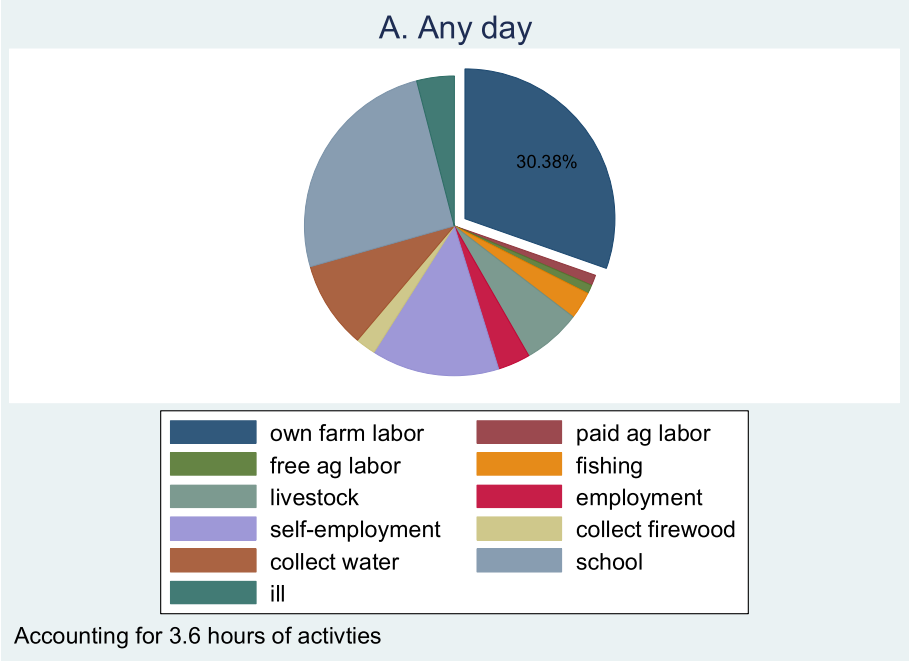
Panel A of Table 4 shows that an average of 1.4 (or roughly 33 percent) fewer household members reported working in farming in the

[13] Throughout the analysis presented here and unless otherwise specified, "plots" refer to plots on which any household member was reported to have worked at any point during the season. This measure of plots depends on the actual incidence of labor (rather than on the stated use of the plots) and so does not include plots held fallow, rented out, and so on, for which no household labor was reported. The analysis is restricted to household members ages 10 or older who reported they worked on any household plot during the season (a "person"). Note that of the 3707 individuals ages 10 and older in the 854 households in our study, 821 reported no agricultural work and are excluded from the analysis. Note also that by this definition of person-plot hours, then, any specific person-plot combination could have a total of zero hours.

[14] The recall NPS households were asked to report the number of days spent performing each of four agricultural activities. They did not provide the specific days on which these activities occurred; so, we do not know if reporting one day in weeding and one day in planting was, in fact, two separate days of work, or a single day in which both of these two activities were performed. To compute total time in hours for the recall NPS group, we chose to compute an upper bound for the number of days by assuming that each activity-day reported was sequential or mutually exclusive, that is, that people did not perform more than one activity on the same day. This choice is supported by the similarity between this measure and the days reported by recall ALT households. It is also supported by the activity patterns of the weekly surveyed households, where we find evidence that agricultural workers overwhelmingly tend to pursue one agricultural activity in a given workday. The typical length of an agricultural workday (roughly four or five hours) as reported across the other three arms of the study is similar to the mean hours per activity in the recall NPS survey, further supporting this interpretation.

[15] Additional comparisons can be made with our survey experiment data and the data from the three waves of the Tanzanian NPS. This is a national panel survey in which sampled households are interviewed once during each survey wave (randomly across 12 months). We can compare the NPS with our weekly data because, in each NPS interview, members were asked if they worked in agriculture, livestock, or fisheries in the previous seven days. This is a broader set of activities than the set here, which is restricted to time spent on the plot. For the NPS subsample of rural households in or near the Mara Region, both participation and hours are significantly higher in the NPS than in our weekly data (results not reported). Hours conditional on working are closer: approximately 26 to 20 h for the NPS and our weekly data, respectively. This suggests that the respondents in the NPS are interpreting the question not literally about hours in the previous seven days, but perhaps are reporting a typical number of hours working. In comparing NPS estimates with estimates obtained in our end-of-season recall modules, we find that both the total days worked on plots and the average hours per working day on plots are roughly the same as in the NPS: 26 days and 4.9 h per working day in the NPS, compared with 29 days and 4.6 and 4.8 h in the two recall designs if analysis is conducted conditional on realized person-plot combinations (not reported).

Panel A



Panel B

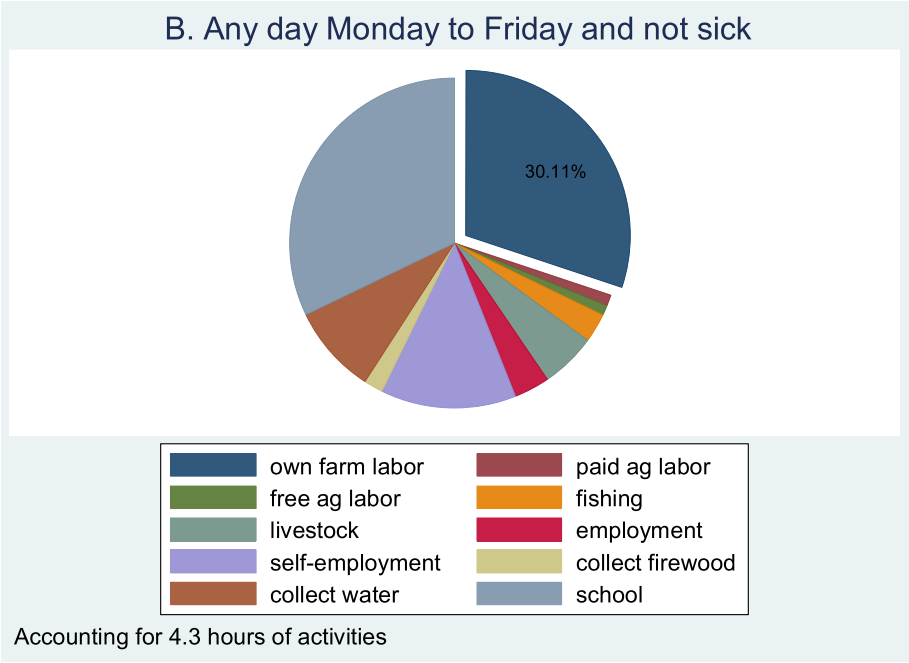
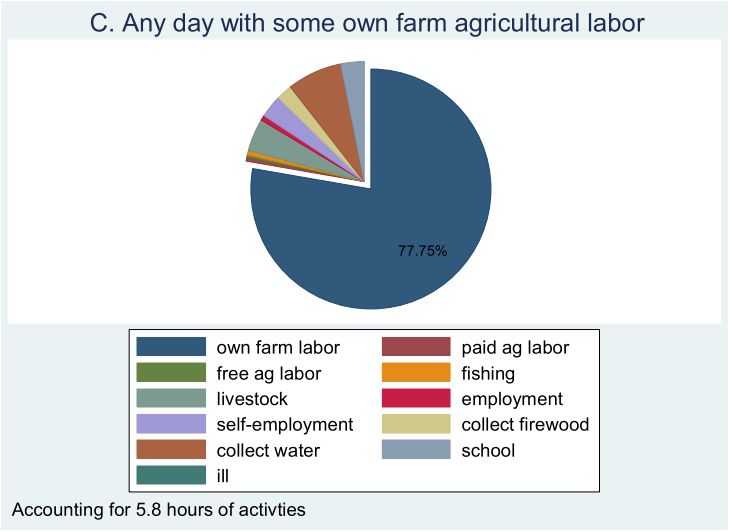


Fig. 1. Activities in an average day. Note: All panels of Fig. 1 are based on Weekly Visit data for household members aged 10 years and over. The data in Panels A and B pertains to all individuals, not just to those individuals reporting agricultural labor at any point in the season.

recall survey. Meanwhile, the number of plots in recall was underestimated by roughly 47 percent, or 2.7 plots (Panel B). For the weekly visits, plot reporting is not fixed at the start of the season. Plots are added after the start of the weekly visits; and some are later dropped (but far fewer). The reasons given for these changes are listed in Appendix Table 1. As was the case in total season hours, the number of people and plots reported as active in agricultural labor is essentially the same between the two weekly survey designs, and between the two recall designs. Importantly, it does not appear that the average number of late-added plots or people in weekly-surveyed households (i.e., the mechanical opportunity these households had to add people and plots over time) can account for the weekly-recall gaps in the reporting of farm-active people and plots.

Panel C



Note: All panels of Figure 1 are based on Weekly Visit data for household members aged 10 years and over. The data in Panels A and B pertains to all individuals, not just to those individuals reporting agricultural labor at any point in the season.

Fig. 1 (*continued*).

Table 3
Means: Total hours and days of agricultural labor reported over season.

	Weekly Visit	Weekly Phone	Recall NPS	Recall ALT
Per person-plot				
Hours	39.5	48.8***	121.3***	146.3***
	(69.4)	(85.2)	(133.8)	(159.3)
Days	9.2	10.7***	25.7***	29.8***
	(14.2)	(14.9)	(24.6)	(29.6)
Weeks	2.5	2.5	N/A	5.7***
	(3.1)	(3.0)		(5.2)
Hours per day worked	4.3	4.5***	4.6***	4.6***
	(1.8)	(2.0)	(1.2)	(1.1)

Note: Mean values which are significantly different from the mean for the Weekly Visit group are denoted as follows: ***p < 0.01, **p < 0.05, *p < 0.1. All of the calculations are restricted to those aged 10 and older who reported having performed agricultural labor at any point in the season, and those plots reporting a positive number of hours of agricultural labor at any point in the season. The calculations are based on all plausible (but not necessarily realized) person-plot combinations per the preceding definition of individuals and plots. "N/A" indicates that the information is not collected in the Recall NPS survey.

Table 4
People and plots Active in household farming.

	Weekly Visit	Weekly Phone	Recall NPS	Recall ALT
A. People per household				
All people	4.9	5.3***	4.7	4.5***
	(2.4)	(2.5)	(2.7)	(1.8)
People working on the farm	4.2	4.3	2.8***	2.8***
	(2.1)	(2.2)	(1.5)	(1.4)
Plots worked per person	3.5	3.5	2.3***	2.4***
	(1.9)	(1.9)	(1.3)	(1.3)
B. Plots per household				
All plots	5.7	5.2	3.0***	3.1***
	(2.5)	(2.4)	(1.6)	(1.6)
Plots cultivated	4.6	4.4	2.4***	2.4***
	(2.2)	(2.0)	(1.3)	(1.3)
People working per plot cultivated	3.2	3.4*	2.7***	2.8***
	(1.8)	(1.9)	(1.4)	(1.4)

Note: Mean values which are significantly different from the mean for the Weekly Visit group are denoted as follows: ***p < 0.01, **p < 0.05, *p < 0.1. All of the calculations in Panel A are restricted to those aged 10 and older. "All plots" refers to all plots reported by the household, including plots which are fallow, rented out, and cultivated (including those owned and rented in). "Plots cultivated" refers to those plots on which agricultural labor was reported as taking place.

4.2. Mechanisms

Why does recall lead to the sort of misreporting described above, and which elements of the farm labor calculation are most vulnerable to recall bias? Below we argue that the need to infer past labor leads to the overstatement of hours worked, and that a lack of salience leads marginal plots and individuals to be under-reported. We also outline several pieces of evidence on heterogeneity in the extent of recall bias, each motivated by insights from the cognitive and behavioral psychology literature.

4.2.1. Over-reporting of hours, days, and weeks worked

4.2.1.1. Failure of recall-and-count and rate-based strategies. If forgetting were the chief mechanism by which recall bias manifests in the hours data, we might expect weekly interviews to yield higher season-total estimates than end-of-season interviews. As the direction of the bias runs counter to this explanation, forgetting is not consistent with our results. In addition, we can rule out recall-and-count strategies which we argue would be reserved for rare and salient events. Weekly data show that people did some agricultural work for an average 11 of the season's 26 weeks and on 46 of the season's roughly 182 days. Over such a long period, recall NPS and recall ALT respondents are unlikely to use recall-and-count strategies in reporting total days and total weeks, respectively. Thus, motivated by the notion, outlined in Menon (1993), that individuals rely on rate-based strategies to report on frequent and non-salient events such as farm labor, we look for evidence of a regular

Table 5
Modal days farmed per week farmed.

Modal days	Frequency (%)	Distribution of days farmed, for a given mode (%)						
		1	2	3	4	5	6	7
1	24.4	**55.7**	14.9	7.8	6.4	5.6	7.3	2.3
2	12.1	17.9	**41.0**	11.4	8.4	8.2	8.5	4.6
3	7.2	14.7	14.7	**33.8**	11.1	11.8	10.0	3.9
4	6.4	11.8	13.8	12.8	**34.7**	11.2	11.8	3.9
5	10.3	12.2	13.0	13.6	11.3	**34.5**	11.7	3.7
6	29.0	9.1	7.6	9.0	11.0	15.4	**41.8**	6.1
7	10.5	6.2	8.7	8.4	9.5	11.1	15.3	**40.9**

Note: This table is based on the data for Weekly Visit individuals aged 10 or over and considering weeks in which some own-household agricultural labor was reported. We do not consider work reported in the baseline, since working patterns cannot be discerned from the data therein. The table can be read as follows. 29.0% of considered individuals have a modal working week of 6 days (in weeks with any own-household agricultural work). 41.8% of their working weeks actually entailed working six days, while 9.1% of their weeks they worked one day.

Table 6
Modal hours farmed per day farmed.

Modal hours	Frequency (%)	Distribution of hours farmed, for a given mode (%)				
		2	3	4	5	6
1–2	5.4	**48.9**	13.4	21.2	6.6	10.0
3	12.5	11.0	**53.4**	20.5	9.0	6.1
4	48.3	4.5	14.8	**57.0**	13.6	10.2
5	15.2	3.2	10.8	25.9	**46.5**	13.6
6+	18.6	3.5	8.9	18.3	16.0	**53.3**

Note: This table is based on the data for Weekly Visit individuals aged 10 or over. We do not consider work reported in the baseline, since working patterns cannot be discerned from the data therein. Less than 2 percent of all observations on hours per day were under 2 h; 7 percent were more than 6 h.

work schedule from which meaningful rates could be constructed. To uncover what, if any, labor patterns exist during the season, we examine the weekly data.[16]

First, we calculate, for each person, the modal days spent farming during those weeks in which there was any farm work. In Table 5, for each mode of days worked, we show the distribution of actual days worked per week across the agricultural season. The distribution of workdays is essentially bimodal: many people generally worked in agriculture once a week (24 percent), while another group worked six times a week (29 percent). Even though farming is the predominant activity in the region, the farming workweek was short, and the majority of people farmed little each week.[17] There is also substantial deviation from the modal work pattern. For example, of those with a modal farm workweek of six days, fewer than half (42 percent) of their weeks entailed six days of work. For these people, 15 percent of their working weeks consisted of five working days, and 9 percent entailed only one working day. The proportion of all workweeks conforming to the people's modal workday (represented by the diagonal in bold in the table) usually represented under half of the weeks, except for mode-1 individuals, who worked one day a week in 56 percent of their working weeks. The proportions of weeks not conforming to the modal work pattern were relatively evenly spread from one to seven working days. From these data, it is clear that even a person's typical workweek is not that typical, and that their work patterns in an atypical week vary widely.

The case with respect to hours per working day, however, is somewhat different. In Table 6, we present the modal number of hours worked per day in farming. Two patterns emerge. First, in contrast to the bimodal days-per-week patterns above, nearly half of farming workdays consist of four hours of work. Second, a larger share of a person's days is spent working the same number of hours as their modal hours. Thus, a larger share of the days worked are on the bolded diagonal here relative to Table 5. There was less variation in the number of hours worked per day than in the number of days worked per week. Inferences based on a typical workday are therefore likely to be more accurate than those based on a typical workweek, consistent with the results presented in Tables 3 and 4.

Together, these results suggest there may be no work pattern to which farmers can reliably refer in constructing a rate-based survey response. This is consistent with reporting in the focus group discussions, which emphasized the variation across the season. Furthermore, we find that the spacing of workdays or working weeks was not consistent over the season, and that the variation in days per week or hours per day observed in Tables 5 and 6 was not driven by seasonality (not reported). For the smallholders in our study, work schedules were both variable (that is, they are different from one week to another) and irregular (that is, there is no systematic or predictable pattern to the variability in work across weeks). Indeed, Table 3 shows that agricultural labor did not take place every day, nor did it even necessarily take place every week. This in turn means that any mental shortcuts or rules of thumb used in inference (for example, “I may not work every day, but I usually work every three days” or “I typically work four days a week”) may produce inaccurate estimates of season-long labor.

In settings such as this, where neither recall-and-count nor rate-based strategies are plausible, how do individuals arrive at the labor figures they report? Several possibilities are raised in the cognitive psychology literature (see, e.g., Das et al. (2012), Godlonton et al. (2016), and de Nicola and Giné (2014) for applications to economic data collected by recall). For instance, people might infer their labor by extrapolating from salient episodes of work, such as the busiest workweek; anchor their inferences on the most recent workweek; or attempt to calculate a total from their knowledge of a rough average. In all cases, the season-wide total is built on the basis of some subset of the season. As a rough exercise to see which subset of the season may be being used as a reference period for respondents' season-wide inferences, we compare the labor reported by end-of-season recall to extrapolations from the weekly data, wherein we scale up one week's worth of hours (alternately, the season-average working week, harvest-average week, peak week, and most recent week) by the 26 weeks in the season. These results are presented in Appendix Table 2. Although none of these scaled-up periods provide a tight approximation of the hours obtained by recall, the totals inferred from the most recent work experiences appear the closest to those obtained by recall, a finding consistent with those in Schwarz and Oyserman (2001) and

[16] For simplicity's sake, these statistics are calculated at the person level (that is, summed across all plots on which each person worked), rather than the person-plot level. If anything, this will understate the degree of irregularity in person-plot working patterns because there is considerable irregularity in the work on a specific plot (see below in this subsection).

[17] This reality has implications for the traditional calculation methodology on labor and labor productivity in agriculture, which tends to assume full-time engagement in farming. Even the recall-based weeks worked and hours worked per day, which we posit are overestimates, are not sufficiently high to support these standard assumptions.

Godlonton et al. (2016).[18]

4.2.1.2. Level of granularity. Another issue may be that survey questions are posed at a level of granularity that is neither intuitive nor intrinsically meaningful to respondents. For example, de Mel et al. (2009) study non-farm enterprise income reporting and find more precision in measures of profits based on a single question asking about profits in the aggregate, compared to those based on adding up many smaller, more granular components of profits. In our context, survey designers may ask about person-plot level labor in order to make plot-level productivity calculations at a later stage, assuming that this sort of granularity comes without a cost in accuracy or ease of recall. However, if farmers tend to think about labor at the person level rather than at the person-plot level, then in an attempt to answer the question as it was posed to them, they may erroneously substitute their person-level estimates for their person-plot level labor. A series of rough comparisons show that such a scenario may indeed be plausible. For instance, the per person hours reported by those surveyed weekly (201 and 228) are closer to the per person-plot hours reported by those surveyed by recall (121 and 146) than they are to what the recall individuals reported at the person level (313 and 389); see Appendix Table 3 for more. Similarly, there is some evidence that recall-surveyed respondents may have reported working on nearly every plot as much as weekly-surveyed individuals report working on any plot: for example, weekly visit individuals report working 46.4 days in total on any plot, while recall ALT individuals report working almost as many days (29.8) at the person-plot level. (For reference, at the person-plot level, a weekly visit individual reports having worked only 9.2 days over the entire season).

Finally, focus groups conducted in the summer of 2016 as a follow-up to the endline survey lend credence to the idea that respondents may be prone to responding at intuitive levels of granularity, irrespective of the format of the question, and often without first mentally adjusting or correcting the response—an act which is cognitively burdensome. To wit, despite being asked to describe typical labor inputs with respect to the size of their actual main plot, respondents often requested to report their answers in terms of 1–2 acre units. Insofar as survey formats are poorly aligned with the way respondents actually think about their work, they will introduce error into labor calculations. In our case, the tendency to conceive of one's work globally rather than on a plot-specific basis, serves to inflate person-plot labor estimates.

4.2.1.3. Cognitive burdens of constructing a response. The mechanisms above suggest that together, irregularity in working patterns and counterintuitive question formats may make constructing a response cognitively burdensome. If this is the case, then individuals with better cognitive skills should be less prone to recall bias. Indeed, in the context of willingness-to-pay experiments, Bergman et al. (2010) found that anchoring bias is mitigated in respondents with higher cognitive skill. In Table 7, we show evidence of precisely this: more highly-educated recall-surveyed individuals were less likely to overstate hours than their less-educated neighbors. The reduction in their overstatement of hours was both large (−23.51 hours, or roughly 25% of the recall premium) and statistically significant. This finding corroborates the idea that the overestimation of hours stems in part from the cognitive burdens associated with inference where irregularity in the working schedule disallows easy rate-based calculations. Furthermore, it carries an important implication for work on agricultural productivity. Namely, it suggests that in studies using recalled labor data, the gains in labor productivity associated with higher human capital may be partly an artefact of education-based differences in labor data accuracy.

[18] In our context, the most recent period coincides with both the peak work period and a particularly culturally and economically salient one, namely, the harvest. For this reason, it is difficult to disentangle the effects of recency from those of work intensity or salience. For instance, if recall NPS and recall ALT household members reported labor based on the work they performed during the last weeks of the season, this could be because the most recent work performed is the easiest to remember, because this is a peak and time-bound work period, or because the work period coincides with the harvest, where work is most salient in terms of income gains.

Table 7
Education interaction regressions: Total hours and days of agricultural labor reported over season.

	(1)	(2)
	Hours Per person-plot	Days Per person-plot
Recall	95.82***	19.00***
	(2.27)	(0.42)
More than primary school	−4.68	−1.32**
	(3.15)	(0.59)
Recall * More than primary school	−23.25***	−4.67***
	(6.46)	(1.20)
r2	0.15	0.16
N	11,542	11,542

Notes: *p < 0.1, **p < 0.05, ***p < 0.01. The sample is those aged 10 and older who reported having performed agricultural labor at any point in the season.

4.2.1.4. Salience of the work performed. Finally, if salience helps in accurate recall, we might expect that the less salient the work, the greater the over-reporting of hours. Thus, we might expect that work done on distant and infrequently visited plots, or work done to cultivate relatively low-labor-intensity crops such as cassava, would result in greater recall bias. Table 8 shows the results where recall assignment is interacted with various plot characteristics. These results broadly confirm the intuition that less-salient plots suffer greater recall bias than do more-salient ones. The fact that the recall bias interacts with plot characteristics can complicate basic stylized facts on agriculture. Our results show, for example, that recall modules will exaggerate hours worked on cassava, and therefore will exaggerate the average labor intensity of cassava cultivation.

4.2.2. Under-reporting of people and plots

4.2.2.1. Forgetting less-salient people and plots. Next, we examine how some of the same mechanisms underlying the over-reporting of hours also drive the under-reporting of people and plots active in agricultural

Table 8
Labor reported with plot characteristics and recall survey interaction.

	(1)	(2)
	Hours per person-plot	Days per person-plot
Plot area		
Recall	91.71*** (2.50)	18.24*** (0.46)
Plot area ≤ 10th percentile	−8.14** (3.94)	−1.87** (0.73)
Recall * Plot area ≤ 10th percentile	−0.85 (5.44)	0.10 (1.01)
Plot area ≥ 90th percentile	36.61*** (3.71)	7.59*** (0.69)
Recall * Plot area ≥ 90th percentile	26.93*** (7.78)	4.04*** (1.45)
r2	0.16	0.17
Plot distance		
Recall	116.7*** (4.223)	22.39*** (0.79)
Plot ≤ 30 min	−4.77* (2.46)	−0.09 (0.46)
Recall * Plot ≤ 30 min	−33.22*** (4.92)	−5.75*** (0.92)
r2	0.15	0.16
Any cassava		
Recall	76.60*** (2.76)	15.21*** (0.51)
Any cassava	−3.97* (2.42)	0.61 (0.45)
Recall * Any cassava	35.17*** (4.42)	6.72*** (0.82)
r2	0.15	0.17
Plot owned by household		
Recall	90.05*** (4.89)	17.96*** (0.91)
Owned	9.79*** (2.33)	2.37*** (0.43)
Recall *Owned	1.95 (5.44)	0.19 (1.01)
r2	0.14	0.16

Notes: *p < 0.1, **p < 0.05, ***p < 0.01. Results from eight regressions. Sample size is 11,542 for each regression. The sample is restricted to those aged 10 and older who reported having performed agricultural labor at any point in the season. The 10th percentile is about 0.08 ha; the 90th percentile is about 0.85 ha. Constant term not shown.

Table 9
Plot characteristic regressions and reporting in recall survey.

	(1)	(2)
	Plot is in recall	Among all cultivated plots, plot is in recall
Plot distance less than or equal to 30 min	0.084***	0.095***
	(0.02)	(0.02)
Plot area ≤ 10th percentile	0.02	0.03
	(0.03)	(0.03)
Plot area ≤ 90th percentile	−0.05	−0.05
	(0.03)	(0.03)
Plot owned	0.09***	0.05**
	(0.02)	(0.02)
Plot owner female	0.003	0.014
	(0.02)	(0.02)
Any Cassava		0.03*
		(0.02)
r2	0.101	0.07
N	3338	2910

Notes: *p < 0.1, **p < 0.05, ***p < 0.01. OLS of recall (=1) or weekly (=0). Sample consists of all plots in the survey. Dummies for missing area and distance included, and as with the constant term, are not shown. The 10th percentile is about 0.08 ha; the 90th percentile is about 0.85 ha.

Table 10
Individual characteristics and reporting of farm work in recall survey.

	Recall
More than primary school	−0.01
	(0.03)
Adult (>19yrs)	0.13***
	(0.02)
Male	0.04**
	(0.02)
Main occupation: farmer	0.06**
	(0.03)
Worked less 30 days	−0.05**
	(0.02)
HH head years of formal education	−0.003
	(0.00)
r2	0.03
N	2679

Notes: *p < 0.1, **p < 0.05, ***p < 0.01. OLS of recall (=1) or weekly (=0). Sample consists of individuals 10 years and older reporting work on the household farm at any point in the season. Constant term not shown.

work. Unlike the case of working time, the direction of the recall bias in the reporting on people and plots active in agriculture means that forgetting is a plausible explanation for the observed gap between weekly surveys and recall surveys.

Appendix Table 1 shows that the primary self-reported reason for adding a plot after baseline was that it had been forgotten during previous visits. Appendix Table 4 shows that plots which were farther, smaller, not owned, and male owned were more likely to be added in later rounds during the weekly surveys. But how do plot profiles compare between recall and weekly surveys? We use the full sample of all listed plots (irrespective of engagement in farm work) and regress the likelihood that they belong to recall-surveyed households against various plot characteristics. Thus, we can get a sense of which types of plots are over-represented in recall relative to in weekly-surveyed households. We present these results in Table 9.

If a plot's salience makes it easier to remember, then we might expect, for instance, more distant plots to be under-represented in recall, while we might expect owned (i.e., as opposed to rented or borrowed) plots to be over-represented. Indeed, this is what we find in Table 9. Here, there is limited evidence that plot size was systematically related to its cultivation. In Table 8 the non-salience of cassava labor was associated with over-reporting of hours in recall, but cassava plots themselves were not systematically forgotten, perhaps because intercropping is common.

We present the results of a similar exercise for people (specifically, household members aged 10 and older who reported working on the farm at any point during the season) in Table 10. We might expect adults, who typically work more frequently and regularly on the farm, to be more easily remembered than children. We find evidence that this is the case: adults are significantly more likely to report working on the farm in recall than in weekly modules. For contrast, there is no evidence of selective recall on measures of education, and only weak evidence on selective recall related to gender. People who worked fewer than 30 days over the season were more likely to be reported in the weekly interviews than in recall. This might be another contributor to the inflation of average hours per person-plot in recall: the average individual reporting agricultural labor in recall is one who is more intensively and regularly engaged in farm work, while in weekly arms of the study, more casual or incidental workers are also captured. Finally, we might expect that people who self-identify as farmers in terms of their stated occupation will be less likely to be forgotten in end-of-season surveys, and this is what we find: in recall, those with the stated occupation of farmer were over-represented in the reporting of farm labor, relative to in the weekly data. Taken as a whole, these results suggest that the non-salience of specific plots and individuals, together with the broader irregularity and inconsistency of farm work patterns, may contribute to the omission in recall of individuals only occasionally engaged in farming.[19]

5. Conclusion

How accurate are data on household farm labor? Our survey experiment finds that recall data collected in the post-harvest period lead to overestimates of the time household members spend on specific plots over the course of the season, in some cases by a factor of 3.7. Recall bias appears to result both from forgetting and from the extrapolation of season-wide labor from erroneous inferences about past labor. Both of these distortions are rooted in the irregular nature of farm-work schedules and practices in our study region. In the absence of a typical work schedule or a typical and consistent level of engagement among workers and on plots, traditional end-of-season recall surveys force respondents into cognitively taxing calculations. These calculations result in labor inferences that appear to be based on recent rather than representative experiences, the omission of members only intermittently engaged in family farm labor, and the exclusion of plots further from the house and, thus, less salient in memory.

This paper makes two contributions to the literature. The first contribution is to the literature on measurement. If our results hold in other settings, then in agriculture-based low-income countries, asking about farm activities 6–12 months after they have ended will lead to exaggerated estimates of the total days and hours household members spend working on their plots and farms. These findings may even hold outside the context of agriculture, for instance, in settings in which some but not other components of the labor calculation face considerable variability (for example, see Dupas et al., 2015).

Clearly, survey designers should tread lightly when asking questions about the frequency of non-salient, irregular events. But what is the alternative? The benchmark weekly visit approach used here is an expensive one that is unlikely to be a realistic prospect at the larger scale necessary for national labor surveys. A result that comes out forcefully in this study is the strong performance of the phone surveys, which show little difference in labor estimates relative to the benchmark weekly visit

[19] Although it may theoretically be possible that the plots that are forgotten are forgotten because they are the prime responsibility of individuals who are themselves likely to have gone unreported, we find little evidence of this. Looking for significant differences in the average number of plots worked per person and the average number of people working per plot, presented in Table 4, we see that both that many different people work on a given plot (an average of 3.2 people in weekly visit households) and that a given person will work on many different plots (an average of 3.5 plots). Thus, it is unlikely that the omission of a single household member would necessarily result in the omission in their plots.

Table 11
Per-household interviewing cost increases

# Interviews	Weekly Visit	Weekly Phone
1	14%	6%
10	139%	54%
20	277%	108%
25	346%	135%
30	416%	162%

Note: The costs are the cost increases in US Dollars, per household, relative to the cost of an LSMS-type (baseline) survey.

survey. Crucially, given the significantly lower transportation costs involved, phone surveys are also, by design, likely to be less expensive to implement than face-to-face high-frequency alternatives, but how much cheaper?

We use the cost data available through our survey experiment to project a scenario whereby an existing household baseline survey adds either short face-to-face surveys or short phone surveys. The results of this costing exercise are presented in Table 11. We assume that all fixed costs related to training and preparation have been subsumed in the baseline interview and focus instead on the increase in the variable costs of conducting 1, 10, 20, 25, or 30 visits or phone calls. Phone calls are much less expensive than in-person visits. The cost of a single round of phone surveys is 6 percent of the cost of the baseline survey. This estimate is close to the 7 percent reported by Dillon (2012). Contacting all respondents 10 times by phone would increase the cost of the survey by 54 percent, while calling all respondents 30 times would increase costs by 162 percent. Our particular experiment required 24 calls to cover the complete agricultural season, but this is highly context-specific, and other surveys may be able to achieve gains in accuracy with fewer points of contact.

These numbers suggest that, in practice, the use of high-frequency phone surveys to collect more reliable labor data may be quite expensive. Nonetheless, it may represent a viable option in surveys that already use phone calls to respondents for other purposes, such as to ensure the continued participation of respondents, to keep track of respondents who relocate, or to collect data requiring high-frequency points of contact or a quick turnaround (Dillon, 2012; Garlick et al., 2015).

One idea that emerges from our study is that although ever more granular reporting may be attractive in that it enables a range of analysis—from plot-level productivity calculations to studies of intra-household allocation—obtaining this level of detail is not costless in terms of accuracy. The appropriate level of aggregation in survey questions will depend both on the purposes to which the data will be set, and the ways in which people think about their work.[20] Accordingly, and given the importance of cognitive burdens in driving mismeasurement in labor data obtained by recall, another approach is to design surveys in ways that minimize these burdens. For instance, where the analytical demands on the data make this possible, questions could be posed in ways that are more intuitive relative to, and better aligned with, the ways farmers actually remember and make inferences about their work.[21] Similarly, data collectors can attempt to shorten the recall period so that labor reporting is likelier to be based on memory than on inference.

Another approach involves managing and correcting for known shortcomings in recall survey data. For instance, by assessing the degree of irregularity in farming practices in the survey context, data collectors will be able to anticipate more effectively whether and the extent to which the resulting labor data will be reliable. They may also use high-frequency surveys such as the ones used in this study, which dramatically shorten the traditional season-long recall period, as an approach to large-scale data collection or as a means to create a consistent adjustment factor that can be applied to past and future recall surveys in the traditional vein. Of course, whether the latter is a reasonable approach to correcting systematic bias in reporting will depend on the specifics of the research context and the degree of variability in these specifics within a given survey group, for instance, the location by region, the crop, the degree of irregularity in farming, the degree of individual responsibility over plots, the prevalence of other types of economic activity, and the uses to which the resulting data will be put. For example, in a similar study conducted in rural Ghana, recall data overestimated the time household members spend on plots by 18 percent (Gaddis et al., 2017). These differences call for attention to the context and characteristics of the population under investigation.

The second contribution of this study is to the debate on the agricultural productivity gap (Gollin et al., 2014). Systematically overestimated measures of the amount of work people carry out on smallholder farms lead to underestimates of labor productivity in agriculture. Furthermore, it is likely that recall-related mis-measurement is correlated with individual, plot and crop characteristics. For example, we find that more highly educated respondents produce less recall bias in reported family farm labor than do their less well educated counterparts. If people with greater levels of human capital are less likely to overestimate their labor in recall, perhaps because they are more well able to cope with the cognitive burdens of remembering and inferring irregular labor, then their higher labor productivity may not be entirely attributable to true differences in productivity driven by skill and education, but, rather, to differences in the quality of labor reporting by level of education. Another example is that farmers assigned to the recall surveys in our study exaggerated time spent farming cassava more than they exaggerated the time spent farming other crops. This would serve to make cassava seem like a relatively more labor-intensive crop than it actually is, purely because its labor inputs are less salient.

Although we draw attention to this one dimension of labor misreporting—namely, that due to recall bias—we acknowledge that there are still other types of labor misreporting that circumscribe our study and others like it. Specifically, we may be operating under a rather narrow concept of farm labor, when the fact is that there is more to farming than what goes on at the plot. For instance, our study may fail to capture the farmer's day in sufficient detail, such as by accounting for the time spent fixing tools, planning for contingencies, negotiating land and labor agreements, and all the other economic and social interactions that are crucial to farm life. Whether the issue is as lofty as fostering structural transformation or as modest as improving data quality, it is clear that a better understanding of the farming context, including the patterns or the lack of patterns in time use, is key.

[20] For instance, in our particular setting, aggregating plot-person hours to the household level, as is done in Appendix Table 3, appears to cancel the competing biases arising from over-reporting at the intensive margin and underreporting at the extensive margin. (Appendix Table 3 reproduces Table 3 in Panel A, and presents statistics at the person level [that is, all labor performed by a given person on any plot] in Panel B, the plot level [that is, all labor performed on a given plot by any person] in Panel C, and the household level in Panel D. Here we can see that the large difference between the weekly and recall surveys virtually disappears in aggregation.). This would mean that even despite recall bias, in our setting, recall data could be acceptable for household-level analysis even if it were unsuitable for, say, plot-level analysis. However, it should be noted that we have no reason to believe that bias would fully or even partially offset by aggregation in other settings.

[21] For instance, it is highly unlikely that individuals are easily able to think about their work in a highly fragmented manner, such as at the person-plot-activity-subseason level. Indeed, survey experiments that test the level—for example, person-plot, person, and so on—at which individuals provide the most accurate labor histories would be a promising area for future research.

Acknowledgements

This study is an output of the "Minding the (Data) Gap: Improving Measurements of Agricultural Productivity through Methodological Validation and Research" project led by the Living Standards Measurement Study team of the World Bank and funded by the U.K. Department for International Development. Additional funding was received from the

IZA/DFID Growth and Labour Markets in Low Income Countries Program (GLM-LIC) under grant agreement GA-C3-RA1-360 and from the World Bank Research Committee. The authors gratefully acknowledge the comments of Doug Gollin and participants at the Centre for the Study of African Economies' Research Workshop, the European Survey Research Association Conference, the Growth and Labour Markets in Low-Income Countries Programme Conferences, the Northeast Universities Development Consortium, the Structural Transformation of African Agriculture and Rural Spaces Conference, and seminar series at the Paris School of Economics, the University of Antwerp, the University of Namur, the University of Washington, and the World Bank. The data were collected onsurveybe and the fieldwork has been expertly implemented by Economic Development Initiatives.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jdeveco.2017.10.005.

References

Anderson Schaffner, Julie, 2000. Employment. In: Grosh, Margaret E., Glewwe, Paul (Eds.), Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Development Study, vol. 1. Washington, DC: World Bank; New York: Oxford University Press, pp. 217–250.
Arthi, Vellore, Fenske, James, 2016. Intra-household labor allocation in colonial Nigeria. Explor. Econ. Hist. 60, 69–92.
Backiny-Yetna, Prospere, Steele, Diane, Djima, Ismael Yacoubou, 2014. The Impact of Household Food Consumption Data Collection Methods on Poverty and Inequality Measures in Niger. Policy Research Working Paper 7090. World Bank, Washington, DC.
Bardasi, Elena, Beegle, Kathleen, Dillon, Andrew, Serneels, Pieter, 2011. Do labor statistics depend on how and to whom the questions are Asked? Results from a survey experiment in Tanzania. World Bank. Econ. Rev. 25 (3), 418–447.
Beegle, Kathleen, Carletto, Calogero, Himelein, Kristen, 2012a. Reliability of recall in agricultural data. J. Dev. Econ. 98 (1), 34–41.
Beegle, Kathleen, Christiaensen, Luc, Dabalen, Andrew, Gaddis, Isis, 2016. Poverty in a Rising Africa. Africa Poverty Report. World Bank, Washington, DC.
Beegle, Kathleen, De Weerdt, Joachim, Friedman, Jed, Gibson, John, 2012b. Methods of household consumption measurement through surveys: experimental results from Tanzania. J. Dev. Econ. 98 (1), 3–18.
Bergman, Oscar, Ellingsen, Tore, Johannesson, Magnus, Svensson, Cicek, 2010. Anchoring and cognitive ability. Econ. Lett. 107 (1), 66–68.
Bound, John, Brown, Charles, Mathiowetz, Nancy, 2001. Measurement error in survey data. In: Heckman, James J., Leamer, Edward (Eds.), Handbook of Econometrics, vol. 5. Elsevier Science, Amsterdam, pp. 3705–3843.
Brown, N.R., Williams, R.L., Barker, E.T., Galambos, N.L., 2007. Estimating frequencies of emotions and actions: a web-based diary study. Appl. Cogn. Psychol. 21, 259–276.
Chen, Shaohua, Ravallion, Martin, 2007. Absolute poverty measures for the developing world, 1981–2004. Proc. Natl. Acad. Sci. 104 (16), 757–762.
Das, Jishnu, Hammer, Jeffrey, Sánchez-Páramo, Carolina, 2012. The impact of recall periods on reported morbidity and health seeking behavior. J. Dev. Econ. 98 (1), 76–88.
Deininger, Klaus, Carletto, Calogero, Savastano, Sara, Muwonge, James, 2011. Can diaries help in improving agricultural production Statistics? Evidence from Uganda. J. Dev. Econ. 98 (1), 42–50.
de Nicola, Francesca, Giné, Xavier, 2014. How accurate are recall Data? Evidence from coastal India. J. Dev. Econ. 106, 52–65.
de Mel, Suresh, McKenzie, David, Woodruff, Christopher, 2009. Measuring microenterprise profits: must we ask how the sausage is made? J. Dev. Econ. 88 (1), 19–31.
De Weerdt, Joachim, Beegle, Kathleen, Friedman, Jed, Gibson, John, 2016. The challenge of measuring hunger through survey. Econ. Dev. Cult. Change 64 (4).
Dillon, Brian, 2012. Using mobile phones to collect panel data in developing countries. J. Int. Dev. 24 (4), 518–527.
Dupas, Pascaline, Robinson, Jonathan, Saavedra, Santiago, 2015. The Daily Grind: Cash Needs, Labor Supply, and Self-control. Unpublished working paper. Stanford University, Stanford, CA.
FAO (Food and Agriculture Organization of the United Nations), 2009. How to feed the world in 2050. In: Issue Brief, High-level Expert Forum, October 12–13, Rome.
Fermont, Anneke, Benson, Todd, 2011. Estimating Yield of Food Crops Grown by Smallholder Farmers: a Review in the Uganda Context. IFPRI Discussion Paper 01097. International Food Policy Research Institute, Washington, DC.
Gaddis, Isis, Oseni, Gbemisola, Palacios-Lopez, Amparo, Pieters, Janneke, 2017. Measuring Farm Labor: Survey Experimental Evidence from Ghana (Unpublished Mimeo).
Garlick, Rob, Orkin, Kate, Quinn, Simon, 2015. Call Me Maybe: Experimental Evidence on Using Mobile Phones to Survey African Microenterprises. Working paper. Duke University, Durham, NC.
Godlonton, Susan, Hernandez, Manuel A., Murphy, Mike, 2016. Anchoring Bias in Recall Data: Evidence from Central America. IFPRI Discussion Paper #01534.
Gollin, Douglas, Lagakos, David, Waugh, Michael E., 2014. The agricultural productivity gap. Q. J. Econ. 129 (2), 939–993.
Irz, Xavier, Lin, Lin, Thirtle, Colin, Wiggins, Steve, 2001. "Agricultural productivity growth and poverty alleviation. Dev. Policy Rev. 19 (4), 449–466.
Ligon, Ethan, Sadoulet, Elisabeth, 2007. Estimating the Effects of Aggregate Agricultural Growth on the Distribution of Expenditures. Background paper for World Development Report 2008. World Bank, Washington, DC.
McCullough, Ellen B., 2016. Labor productivity and employment gaps in sub-Saharann Africa. Food Policy 67, 133–152.
Menon, Geeta, 1993. The effects of accessibility of information in memory on judgments of behavioral frequencies. J. Consum. Res. 20 (3), 431–440.
Olinto, Pedro, Beegle, Kathleen, Sobrado, Carlos, Uematsu, Hiroki, 2013. The State of the Poor: where Are the Poor, where Is Extreme Poverty Harder to End, and what Is the Current Profile of the World's Poor? Economic Premise 125 World Bank, Washington, DC.
Reardon, Tom, Glewwe, Paul, 2000. Agriculture. In: Grosh, Margaret E., Glewwe, Paul (Eds.), Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Development Study, vol. 2. Washington, DC: World Bank; New York: Oxford University Press, pp. 139–182.
Ross, Michael, Conway, Michael, 1986. Remembering One's own past: the reconstruction of personal histories. In: Sorrentino, Richard M., Tory Higgins, E. (Eds.), Handbook of Motivation and Cognition: Foundations of Social Behavior, vol. 1. Guilford Press, New York, pp. 122–144.
Schwarz, Norbert, Oyserman, Daphna, 2001. Asking questions about behavior: cognition, communication, and questionnaire construction. Am. J. Eval. 22 (2), 127–160.
Sudman, Seymour, Bradburn, Norman, 1973. Effects of time and memory factors on response in surveys. J. Am. Stat. Assoc. 68 (344), 805–815.
World Bank, 2014. Final Report. Report 90434. Vol. 2 of Tanzania: Productive Jobs Wanted. Country Economic Memorandum. World Bank, Washington, DC.