How does recall bias in farm labor impact separability tests?

Bailey Peterson-Wilhelm and Benjamin Schwab*

June 24, 2024

Abstract

In the agricultural household literature, empirical tests of separability between production and consumption decisions commonly exploit theoretical predictions of household labor allocation. Many of these studies rely on data that asks respondents to recall labor usage over the entire growing season. Two recent field experiments in Tanzania and Ghana show that such labor use data, collected at the end of the growing season, is a systematically unreliable measure of actual labor allocation. In this study, we examine how inaccurate measures of labor influence the reliability of market failure tests based on separability. In Ghana, we find no statistical evidence that recall bias influences the reliability of the separability test. In Tanzania, we find that recall bias increases the probability that such tests fail to reject separability. Thus, we find partial evidence that classic tests based on typical household survey labor data may erroneously conclude that markets are adequately functioning.

Keywords: Separability, Recall bias, Ghana, Tanzania, Labor, Agricultural Household

Acknowledgement The authors acknowledge funding from National Science Foundation grant #1828571. We thank Brian Dillon for helpful comments. All errors are our own.

*Schwab: Associate Professor, Department of Agricultural Economics, Kansas State University, benschwab@ksu.edu; Peterson-Wilhelm: PhD Student, Department of Agricultural Economics, Kansas State University, bailey27@ksu.edu

1 Introduction

In the classic agricultural household model, farm production decisions are nested within the household's optimal consumption problem. The joint structure implies several threats to the assumption that agricultural production decisions are made separately from the household's consumption decision. The separation assumption requires environments with complete markets. When a household cannot access certain input or output markets for example, then separation is unlikely to hold, and its production decisions will not be consistent with a profit-maximizing firm.

Benjamin (1992) devised the classic empirical test of separability by focusing on the household labor allocation decision. If the size and nature of the family household labor endowment statistically influences its total labor used in agricultural production, separability does not hold (Benjamin, 1992). Recent work expanding the separability test continues to rely on measures of labor use (Dillon and Barrett, 2017; Dillon, Brummund, and Mwabu, 2019).

These studies typically rely on data collected via a survey conducted at the end of the growing season. The agricultural modules of such surveys are the foundation of often used data sets like the World Bank Living Standards Measurement Study and Integrated Surveys on Agriculture (LSMS-ISA). A growing literature has increasingly focused on measurement error in such agricultural data (Carletto, Dillon, and Zezza, 2021). In particular, two recent field experiments, one in Tanzania and one in Ghana, have revealed systematic errors in 'one-shot' survey-based labor measures. These experiments find meaningful disparities between labor measures obtained via end-of-season surveys and intensive weekly surveys.

In Ghana, Gaddis et al. (2021) randomized the data collection procedure by assigning some agricultural households to weekly data collection and others to the traditional end of growing season method. They find 10% overstatement of farm labor on a per person per plot basis, driven by 'listing bias', or systematic failure to report on more marginal plots (and workers). In failing to accurately recall all plots and workers, such bias underestimates total

household labor due to the omission of certain plots and laborers, and thus tends to inflate per unit measures of labor productivity. The bias is referred to as recall bias because it is tied to the inaccurate ability to recall plots, workers, and labor used by the household.

The Gaddis et al. (2021) results complement a similar study by Arthi et al. (2018), which compares weekly surveys to two different end of season methods in Tanzania. The Tanzanian experiment similarly finds substantial recall bias, with four times as much labor per person per plot reported by the recall group as compared to the group surveyed weekly. As in Ghana, omission of plots appears to play a substantial role in the discrepancies between intensive and end-of-season measures of labor use.

Both experiments identify consistent sources of non-classical mismeasurement of household labor supply from end-of-season recall. Households fail to recall some plots, and thus do not account for labor on them. They also exaggerate the quantity of labor on the plots they do list. However, the implications for household level aggregates of these competing sources of error differ between the two settings.

In Tanzania, the two sources of bias essentially even out. Arthi et al. (2018) notes that household level aggregation "appears to cancel the competing biases arising from over-reporting at the intensive margin and underreporting at the extensive margin." In contrast, in Ghana, the depressing impact of listing bias dominates. The traditional one-shot surveys underestimated total household labor supply, despite the fact that the ommitted surveys and plots were relatively marginal. Overall, then, it is unclear how further aggregating household labor demand will be impacted by these sources of mismeasurement.

Thus far, the literature on measurement error in agricultural household has focused primarily on issues of productivity. McCullough (2017) notes that problems arising from lack of attention to the distinct measurement issues in household agricultural data have led to overestimation of productivity gaps in the agricultural sector and gains from reallocation of resources. Similarly, both Gaddis et al. (2021) and Arthi et al. (2018) focus on the implications of labor mismeasurement for understanding smallholder agricultural productivity.

However, neither study considers the potential influence of recall bias for empirical separability tests. Conceptually, the direction and magnitude of the effects are ambiguous.

Listing bias arising from a failure to list household members may obscure a true link to overall labor demand, thus causing an underestimation of the key test statistic utilized to reject the null hypothesis of separable labor allocation decisions. However, if marginal plots are relatively less likely to utilize household labor even if separability holds, traditional estimates may overestimate the test statistic and thus falsely reject separability.

Further complicating predictions, neither Gaddis et al. (2021) nor Arthi et al. (2018) describe how recall bias impacts the measurement of *hired* labor. Because empirical separability tests are based on total labor demand (i.e. hired and household labor), hired labor decisions are critical. Indeed, assessing the 'missingness' of the labor factor market is a key goal in the empirical literature.

Measurement error in hired labor may not follow the same pattern as household labor. The necessity of making explicit payments, for example, may increase the salience of hired labor allocations and reduce the types of cognitive and accounting biases that plague end-of-season surveys. Further, the propensity to hire labor may not be uniform across all plots, particularly the 'marginal' plots that Gaddis et al. (2021) and Arthi et al. (2018) find most likely to be omitted.

In this paper, we utilize the data from the original Gaddis et al. (2021) and Arthi et al. (2018) in Ghana and Tanzania to determine how recall and listing bias have impacted the substantial empirical literature on separability tests.

For the first time, we assess how hired labor is impacted by the forms of recall and listing bias identified in the original experiments. The setting in Ghana, where most households hire labor, and Tanzania, where hired labor is rare, provide contrasting contexts to assess how mismeasurement impacts hiring data.

Including hired labor allows us to determine the influence of recall error on total household labor demand, not only household labor supply. We then use that data to estimate the classic

labor separability tests using the Ghana and Tanzania data, and determine how recall bias impacts the validity of the tests common in the literature.

In Ghana, where hired labor is more common, we find little evidence that hired labor suffers from recall bias at either the household aggregate level or on a per acre basis. However, total household labor demand remains significantly underestimated by recall households as a result of the listing bias from household labor supply. In Tanzania, where hired labor is more rare, recall households significantly underestimate the amount of hired labor used.

In both Ghana and Tanzania, we strongly reject the null of separability for both the end-of-season and weekly data collection arms. Despite the large discrepancy between aggregate labor demand as measured via the two methods in Ghana, we find recall bias has no effect on the empirical test of separability. In contrast, in Tanzania, we find that the test statistic obtained from end-of-season data collection is underestimated relative to the weekly visit group.

Thus, in one context, recall bias increases the probability that the classic Benjamin (1992) test fails to reject separability and that tests based on such data are more likely to erroneously conclude that markets are adequately functioning. The finding that labor measurement problems may lead to underrejection of separation confirms speculation in LaFave and Thomas (2016). Unlike LaFave and Thomas (2016), however, our analysis is able to specifically account for the role of non-classical measurement error and the importance of listing bias. Thus, we demonstrate a plausible explanation for the failure to reject separability in the Indonesian data analyzed in Benjamin (1992).

Concerns that measurement error can bias separation tests have been addressed in the recent literature through the use of panel data analysis, which should account for household level, time-invariant errors. However, the form of recall bias illustrated by the experiments in Ghana and Tanzania suggest key sources of agricultural labor mismeasurement vary across time. Listing bias, for example is a function of both invariant and variant household characteristics. For the same household, plots or workers that may be marginal and omitted in one

season may be more salient in another due to changing agroecological characteristics (e.g. rainfall), market conditions (e.g. price changes for crops on different plots), or household member characteristics (e.g. a child aging and become more productive). Thus, while the lack of panel data in the measurement experiments under study here constrain our ability to directly test the impact of labor mismeasurement in panel settings, the results of our analysis provides insight into the potential impacts of recall error that are not directly mitigated by estimating household fixed effects models.

While the analysis here focuses on the implications of agricultural labor mismeasurement, our research connects to a growing literature on the influence of non-classical measurement error in agricultural data.[1] That literature has established the pervasiveness of systematic non-classical measurement error from traditional survey reporting of land size (Carletto, Gourlay, and Winters, 2015), input use (Abay, 2020) and production (Lobell et al., 2020). Revised analyses that account for these errors have overturned conventional wisdom on the existence of the inverse-scale productivity relationship (Gourlay, Kilic, and Lobell, 2019) and the degree of allocative inefficiency in East African agriculture (Gollin and Udry, 2021). In contrast, our findings indicate that accounting for labor measurement error should increase our confidence in the recent literature that rejects rural market completeness (Dillon, Brummund, and Mwabu, 2019; Dillon and Barrett, 2017; LaFave and Thomas, 2016). Because failure to reject separability is interpreted as a positive sign of economic integration, our work cautions that the reliability of the underlying labor use data can contribute to such false negatives. Similar to Abay, Bevis, and Barrett (2021) and Aragón, Restuccia, and Rud (2022), the contrasting implications of measurement error for aggregate household labor, hired labor and separability tests in Ghana and Tanzania show that empirical tests are sensitive to the specific mechanism and form of recall error.

[1]The agricultural measurement literature is related to a larger body of work examining measurement issues arising from household surveys more broadly, particularly for measures of consumption and poverty (see, for example, Abate et al. (2023, 2022); Abay et al. (2021); De Weerdt, Gibson, and Beegle (2020); Kilic and Sohnesen (2019); Zezza et al. (2017); Friedman et al. (2017); Gibson et al. (2015); Beegle et al. (2012) and Beaman and Dillon (2012)).

2 Research Design and Data

Our data are drawn from the original data measurement experiments carried out by Gaddis et al. (2021) and Arthi et al. (2018) in Ghana and Tanzania, respectively. We discuss each experiment in the following sections.

2.1 Ghana Experiment

The data used for analysis were collected in the Ashanti and Brong Ahafo regions of Ghana during the 2015-2016 agricultural growing season. The 720 surveyed agricultural households from 20 villages were then randomly assigned to one of three data measurement groups: traditional end of season data collection (control), weekly visits by enumerators, and weekly phone interviews. Gaddis et al. (2021) excludes the phone group from the main analysis because the phone group was included as a proof-of-concept. For completeness, we include the data from this group here. For both an in-person visit and a phone survey the recall window is significantly narrowed, which is the main mechanism behind reducing recall bias. However, we recognize that recent literature has documented that quality of data varies between in-person and phone surveys (Abate et al., 2023; Jeong et al., 2023; Anderson et al., 2024). Specific to agricultural data collection, Anderson et al.(2024) find that phone responses have a greater variance than in-person responses resulting in less precise estimates. By including the phone group, we are reducing recall bias, but we could be introducing other forms of bias associated with phone surveys. Recognizing this possibility, we take steps to ensure our results are robust to the exclusion of the less precise phone survey data, as discussed in the results section.

All groups were given an in-person baseline survey before the growing season began. The weekly visits group was surveyed weekly by an in-person enumerator. The weekly phone group was given a similar weekly survey by phone call on a phone provided to households. These regular follow-up surveys were kept short, to prevent respondent fatigue (Appendix

Table 2). At the end of the growing season, all groups were given an endline in-person survey.

To maintain consistency, the data was cleaned similarly to Gaddis et al. (2021). Households in all treatment arms that did not complete the endline survey were removed from the sample. Households in both treatment groups that did not complete at least two thirds, or 16, of the weekly surveys were dropped to avoid introducing recall bias into the treatment groups. Child labor is not the focus of this study, so labor of children under 10 years old at the baseline survey was dropped. Household and hired labor hours across all treatment arms were winsorized at the top one percent at the person-plot level. Households that reported using zero labor hours over the season were removed from the sample in addition to households that did not report having any agricultural land. The number of households dropped at each stage is reported in Appendix Table 1. The table highlights that a similar number of households were dropped in each group, alleviating concerns about attrition induced bias.

Aggregate labor demand for each household was constructed by combining household labor and hired labor. The consideration of hired labor is an extension beyond the scope of Gaddis et al. (2021). For both treatment groups, household labor includes labor hours reported at the baseline, weekly hours, and hours reported at the endline. In the baseline survey households were asked to report hours worked by person and plot prior to the initial baseline survey. The weekly survey asked the number of hours worked by person and plot since the last survey. Finally, the endline survey asked the number of hours worked by person and plot since the last weekly visit. Hired labor also includes baseline, weekly, and endline hours reported; however, the questions were asked on the plot level rather than the person-plot level. Control group household labor includes only the hours reported in the endline, where the households were asked to report the number of days spent on each of four activities over the entire growing season by person and plot and the typical day length by activity. Similarly, hired labor for the control group was collected by asking total days spent on each activity by plot for the entire season.

2.2 Tanzania Experiment

The Tanzania data used for this analysis were collected during the 2014 rainy season in the rural Mara Region and were drawn from Arthi et al. (2018). The sample consists of 854 agricultural households spanning 18 enumeration areas, or villages. Within each village the households were randomized to one of four survey arms: weekly visit, weekly phone, traditional recall, and alternative recall. The weekly visit and weekly phone groups were given an in-person baseline survey, weekly surveys, and an in-person endline survey. As discussed in the Ghana Experiment section, we do not expect visit and phone surveys to be equivalent in terms of data quality. However, both weekly methods should reduce recall bias via a similar mechanism. Further, robustness checks are used to ensure that data quality concerns from phone surveys do not drive the results.

The recall groups were only surveyed once at the end of the growing season. For the traditional recall group, the labor module asked households the total number of days spent on each of four agricultural tasks throughout the growing season and the typical numbers of hours spent per day for each task, which is identical to the labor module used previously in Tanzania and very closely mirrors the labor module for the Ghana recall group. The labor module for the alternative recall group instead asks the total weeks worked over the entire growing season, typical number of days worked per week, and typical hours worked per day, without specifying labor activity. To maintain consistency, the same Ghana cleaning rules were applied to the Tanzania data. In order for the Tanzania treatment groups to complete at least two-thirds of the surveys, households must have completed 19 weekly surveys. A similar number of households were removed from the visit and phone groups for not meeting this requirement, as reflected in Appendix Table 1.

Aggregate labor demand includes both household agricultural labor and hired agricultural labor. Arthi et al. (2018) did not include hired labor in their analysis, so the effect of recall bias in hired labor has been previously unexplored. The revisit and phone treatment groups include household and hired labor reported at the baseline survey, weekly surveys,

and endline survey. The traditional and alternative recall groups only include hired and household labor reported in the endline survey.

2.3 Summary Statistics

Household characteristic summary statistics are included in Table 1. We show that the household characteristics that may influence separability are well-balanced between groups. In both Ghana and Tanzania, the visit and phone group household size diverge from the traditional recall group, respectively. Like in the original Gaddis et al. (2021) paper, we argue that the difference in household size is a function of the treatment, rather than a random imbalance. The visit and phone groups both have many opportunities to update their household roster, meaning the visit and phone group household roster is more likely to be complete. We see the same difference in the average number of plots per household but to a much higher degree, which is discussed below. Compared to plots, which may be distant or marginally used, a household member is less likely to be forgotten which explains the lesser statistical difference when compared to the difference noted in the average number of plots. Other household characteristics that influence separability are well-balanced across groups, which gives confidence that any separability implications are due to the treatment, rather than group imbalances.

Summary statistics describing the labor data are presented in Table 2. Household labor is averaged on per worker, per active plot, per acre, and per household levels for all treatment arms in Ghana and Tanzania. Hired labor and aggregate labor demand are averaged on the per active plot, per acre, and per household level for all groups in Ghana and Tanzania. For tests of differences in means, in Ghana and in Tanzania the reference group is the traditional recall group.

In Ghana, the dominant effect of listing bias–manifested in the recall group omitting more marginal plots–is apparent in the treatment arm differences for household labor across the levels of aggregation. At the per person level, the visit group is not statistically different

from the recall group. However, at every other level the visit group (and phone group) report significantly more household labor hours than the recall group. The average number of plots for visit and phone groups are also statistically larger than the recall group. The average number of acres per household is smaller, but not significantly so, reflecting the fact that the omitted plots tend to be smaller and more marginal. The recall group's lagging household-level totals indicate that by forgetting to list marginal workers and their labor, in aggregate the recall group is under-reporting household labor.

In contrast, in Tanzania, visit and phone groups are statistically different from the traditional recall group in granular measures but the difference dissipates as the labor hours are aggregated. The two treatment groups also report significantly more acres and significantly more plots compared to the two control groups. The statistical difference at the per person and per plot level combined with the large difference in plots and acreage, could indicate that the traditional recall group is reporting an accurate number of household labor hours in total but are incorrectly allocating the labor hours. Since the number of plots is lower for the recall group, they are attributing household labor completed on unlisted plots to the plots that are listed. It is unclear whether the downward bias (from failing to list plots) evenly cancels out the upward bias (from overreporting labor hours on listed plots) due to chance or from some heuristic process common to farmers in this setting.

The impact of recall bias on hired labor also differs in both countries.[2] As hired labor requires payment, it is possible that respondents are able to more accurately recall hired labor hours than household labor. In Ghana, where hired labor is common, there is little evidence that hired labor is affected by recall bias. From a mental accounting perspective, households may more accurately recall the amount of hired labor used because of the transfer of funds. At the per-plot level, only the phone group, which suffers from previously noted data quality concerns, differs meaningfully in measures of hired labor. For all other levels of aggregation, hired labor measures are strikingly similar across treatment arms. Hired labor

[2]Beegle, Carletto, and Himelein (2012) exploit the variation in time between survey dates and harvest dates and also find the effect of recall bias on hired labor varies across several African countries.

is more likely to be used on high-quality, salient plots that are less affected by listing bias, which is consistent with the lack of hired labor mismeasurement.

In Tanzania, however, recall bias appears to result in an underestimation of hired labor at the household level. On a per-household basis, the visit group reported nearly 50 percent more hired labor than the traditional recall group, with the difference significant at the 10 percent level. However, the low utilization of hired labor in Tanzania means that the discrepancy accounts for only 2 percent of total labor demand.

Our summary statistics confirm and expand upon the findings of Gaddis et al. (2021) and Arthi et al. (2018). While listing bias and recall bias clearly influence the accuracy of labor modules, the question of how these biases influence common tests of separability remains unanswered.

3 Econometric Specification

Using the survey data on both family and hired labor, we calculate total household labor demand for the agricultural season among all the treatment groups. We then replicate the classic separability tests, as used in Dillon and Barrett (2017), where log total labor demand is regressed on the log of household size, total acres, and household characteristics. The selected household characteristics are the head of the household gender, equal to one if male, share of the household that is an adult female, and share of the household that is an adult male. The share of the household that is children is the omitted category. Our basic regression specification takes the following form:

$$ln(labor_i) = \beta_0 + \beta_1 ln(HHSize_i) + \beta_2 ln(acre) + \beta_3 HeadGender + \beta_4 ShareF_i + \beta_5 ShareM_i + u_i \quad (1)$$

In the standard analysis, if $\hat{\beta}_1 > 0$, the null of separability is rejected. In (1), that would be the equivalent of rejecting the null for the end-of-season data collection. To extend the

base regression specification in (1), we include measures of treatment status ($Treated_i$) and its interaction with the log of household size. The variable $Treated_i$ is equal to one if the household is in the visit or phone group and zero otherwise.

$$\begin{aligned} ln(labor_i) = \beta_0 + \beta_1 ln(HHSize_i) + \beta_2 ln(acre) + \beta_3 HeadGender + \\ \beta_4 ShareF_i + \beta_5 ShareM_i + \beta_6 Treat_i + \beta_7 (Treat_i * ln(HHSize_i)) + u_i \end{aligned} \tag{2}$$

As noted in 2, the weekly surveying structure influences plots and acres in addition to the total labor demand. Therefore, we include an interaction between the treatment status and $ln(acres)$ as well such that:

$$\begin{aligned} ln(labor_i) = \beta_0 + \beta_1 ln(HHSize_i) + \beta_2 ln(acre) + \beta_3 HeadGender + \beta_4 ShareF_i + \\ \beta_5 ShareM_i + \beta_6 Treat_i + \beta_7 (Treat_i * ln(HHSize_i)) + \beta_8 (Treat_i * ln(acre)) + u_i \end{aligned} \tag{3}$$

To determine if the weekly visit treatment influences the standard separability test, we focus on β_7. Estimating $\hat{\beta}_7 > 0$ implies that the relationship between household labor demand and household labor composition is underestimated in typical surveys, and that separability tests are susceptible to falsely under-rejecting the null of separability because of labor recall bias. However, $\hat{\beta}_7 < 0$ would imply the opposite, and that the literature showing strong rejections of separability across, for example, Sub-Saharan Africa (SSA), could potentially be an artifact of non-classical measurement error in labor data.

LaFave and Thomas (2016) argue that separability tests based solely on cross-sectional data potentially suffer from endogeneity concerns. Such concerns threaten consistency in estimating $\hat{\beta}_1$. However, identification of β_7 relies solely on the random allocation of the measurement treatment (i.e. recall length).

We also attempt to empirically verify the extent to which enodgeneity concerns may affect the key parameters by estimating less parsimonious versions of equation (2) that include additional control variables likely to mitigate the omitted variable bias in $\hat{\beta}_1$. As more

potentially correlated variables are removed from the error term and included as controls, instability in the estimates could indicate traditional endogeneity concerns affect the main coefficient estimates. We thus estimate the following equation, adding variables consistent with the specification in Dillon and Barrett (2017):

(4)

$$ln(labor_i) = \beta_0 + \beta_1 ln(HHSize_i) + \beta_2 ln(acre) + \beta_3 HeadGender + \beta_4 ShareF_i + \beta_5 ShareM_i + \beta_6 Treat_i + \beta_7 (Treat_i * ln(HHSize_i)) + \beta_8 (Treat_i * ln(acre)) + u_i + \delta X_{iv} + \theta_v + u_{iv}$$

In (4), X is a vector of additional control variables and village fixed effects (θ). The additional control variables include measures of soil quality and soil health, the maximum household education, and off-farm employment.

4 Results

4.1 Ghana Results

We first analyze the impact of recall bias on separability tests in Ghana by pooling the phone and visit treatment groups and estimating equation (1) (Table 3). Separability is rejected regardless of model specification. Household size significantly increases labor demand (p <0.01). Other household characteristics, such as household family structure, significantly contribute to labor demand (p <0.01), providing further evidence that separability does not hold. Treatment is positive and significant in both the base and extended model (p <0.01), which is consistent with the conclusion from Table 1 that both visit and phone treatment groups report more labor than the recall group.

The interaction of treatment with household size is close to zero and not significant at traditional significance levels in both model specifications. The null effect of the treatment household size interaction indicates that despite the evidence of substantial recall bias at

the household level in traditional labor measurement methods, that bias did not affect the separability test in this context.

Since there was some evidence that plot listings might be susceptible to bias, we considered the treatment effect on acreage in Table 3. Across all of the model specifications acreage significantly contributes to labor demand (p <0.01). Unsurprisingly, as area increases, the demand for labor also increases. The treatment area interaction is also statistically significant (p <0.01); however, it is negative. The negative interaction term indicates that as plots are more accurately accounted for, in this case through additional opportunities to add plots to the plot roster, the impact of acreage on labor demand is reduced by roughly half.

The differences in labor demand across treatment groups and concerns about phone survey data quality prompted us to look closer at the differential impact of each treatment. Table 4 considers the base regression for each treatment independently and then again with separate indicator variables and interaction terms for each treatment.

We find that the results are largely unchanged from the pooled results presented in Table 3. The coefficient on the log of household size remains large, positive, and statistically significant (p <0.01). The treatment effect is still positive and statistically significant (p <0.01) for both treatments in each of the specifications, though the phone treatment effect exceeds the visit treatment effect. The larger phone treatment effect is consistent with the summary statistics in Table 1, where the phone group reported higher labor demand. Once again, each of the treatment household size interaction terms are insignificant. The relative comparability between the results of the visit group and phone group separately indicates that while the data quality might vary, that difference is not the driver of our results. In Ghana, recall and listing bias do not significantly affect the reliability of separability tests.

4.2 Tanzania Results

As in Ghana, we first pool the phone and visit treatment to estimate equation (1). For the pooled results in Tanzania, we omit the alternative recall group altogether, so the pooled

weekly visit treatments are estimated against the traditional recall counterfactual only. The resulting estimates in Table 5 include the base and extended model for Tanzania.

Across all specifications, separability is rejected. The household size coefficient is large, positive and significant (p <0.01). The rejection of separability is further indicated, as other household characteristics also significantly contribute to labor demand. In Tanzania, a male head of household increases labor demand (p <0.1) and the household structure also significantly influences labor demand (p <0.05).

In contrast to the Ghana results, the treatment interaction with household size is positive and significant (p <0.05). The positive and significant interaction between treatment and household size indicates that the impact of household size on labor demand is an underestimate. Thus, recall bias makes it more likely to falsely fail to reject separability.

In Tanzania, the treatment indicator variable is negative and significant across specifications (p <0.05). However, given an average household size of 6.55, the independent effect of correcting for recall bias is near zero. The near zero total effect of the weekly survey treatment is consistent with the summary statistics from Table 1, which reveal little difference in labor demand between treatment and control households at the aggregated household level.

The positive and significant effect of acreage on labor demand (p <0.01) is consistent with the Ghana results. As expected, a larger agricultural area would require additional labor. Differing from Ghana, the interaction of acreage and treatment is insignificant in Tanzania. The insignificance of the interaction term is consistent with the hypothesis that households were not omitting labor, but instead mis-attributing labor from unlisted plots to listed plots.

In Table 6, the treatments are again considered separately. Treatment and the treatment interaction are insignificant for the phone group. However, the weekly visit treatment negatively and significantly affects labor demand, a similar finding to that in Table 5. Crucially, the interaction term between the weekly visit treatment and household size is significant and positive (p <0.01). The positive effect of the interaction term indicates that the effect of

household size is even larger when recall bias is reduced.

When the visit group treatment is considered alone, reducing recall bias increases the effect of household size on labor demand by roughly 40 percent. The sign of the treatment effect coefficient for the visit only and phone only specifications is consistent, though the phone group estimate is smaller and less precise. Thus, the visit group is largely driving the treatment effect in the pooled sample. Since the visit group is well balanced when compared to the traditional recall group in terms of household characteristics, the treatment effects do not appear to be driven by the attrition and small imbalances from the phone survey.

When we consider the alternative baseline group as an additional treatment group rather than a control group, we see it behaves similarly to the visit treatment group. The treatment coefficient is negative and significant (p <0.1) and the interaction between treatment and household size is positive and significant (p <0.05) and of similar magnitude. The treatment model in Table 6 also highlights the heterogeneous impact of the different treatment types.

The estimates in Table 6 indicate that when recall bias is reduced in Tanzania, separability is more strongly rejected. The presence of some significant relationship between the reduction of bias and separability indicates that, at least in Tanzania, recall bias might lead to under-rejection in tests of separability in the literature.

4.3 Robustness Checks

In order to test the robustness of our results, we add a series of additional controls to our existing preferred model in a step-wise procedure. If endogeneity problems arising from omitted variables threaten the validity of the separation tests, the coefficients on the key variables (household size and its interaction with treatment) should be sensitive to models that include more detailed control variables. Taking the estimates in Table 3 & 5 as a base, we estimate equation (4) by gradually adding the additional controls, beginning with the addition of village level fixed effects, captured using binary enumeration area variables. The first additional set of controls include soil quality and soil type. Next, controls for household

education were included. For the presented specification, maximum household education consists of binary variables for the highest level of education by any household member. The final set of controls are measurements of off-farm employment: the number of non-agricultural enterprises, average monthly income of household non-agricultural enterprises, and the share of the household employed off-farm.

The results from the step-wise addition of control variables are presented in Table 7 for the full sample. Table 8 includes the results by treatment arm with all of the selected control variables included. Results are robust to the addition of these additional variables. In both Ghana and Tanzania for the full data set and treatment arm subsets, the point estimates for both the measurement treatment effect estimate (β_7) and primary separability test (β_1) remain largely unchanged as control variables are included.

The stability of the estimates to household education and agricultural covariates suggests that the lack of panel data, which is commonly used to control for idiosyncratic household labor demand shifters, is unlikely to be driving the results or adversely affecting the identification of our main variable of interest, β_7.

5 Concluding Remarks

The evaluation of the complete markets assumption in rural settings has long relied on labor use data collected via surveys of agricultural households. These surveys typically occur once, after harvest, and ask respondents to recall data over the previous growing season. A recent experimental literature has documented that this 'one-shot' method leads to mismeasurement. Generally, compared to those surveyed frequently, households do not list all the laborers or plots that use household labor, and overstate the amount of labor expended on those plots they do list, while simultaneously understating total labor usage.

The experimental literature on labor measurement has focused primarily on the implications of survey recall bias for measures of agricultural productivity. As a result, these studies

analyze household labor supply, not total labor demand, which includes hired labor.

In this paper, we reanalyze data from labor measurement field experiments in Ghana and Tanzania. We examine the consequences of labor mismeasurement for estimates of total labor demand, which includes hired labor, and for empirical tests of separation. Such separability tests estimate the relationship between household composition and total labor demand, and are thus highly dependent on household labor data.

We find that while per plot and per acre measures of hired labor are not generally affected by mismeasurement, recall bias caused households in Tanzania, but not Ghana, to understate the total amount of hired labor employed during the growing season. In both Tanzania and Ghana, the null hypothesis of separability is strongly rejected regardless of data collection frequency. In Ghana, recall bias has no impact on the reliability of the separability test. However, in Tanzania, recall bias causes an underestimation of the separation test statistic. Thus, we demonstrate the possibility that the long empirical literature relying on separation tests may suffer from under-rejection bias, and thus overestimates market completeness.

To better contextualize the role of measurement error in the implementation of tests of separation, we compare the findings here to similar tests in Dillon and Barrett (2017) for five African countries. Using a nearly identical specification, their estimate of β_1 (.588) in Tanzania is very similar to our estimates for the control sample that received the traditional recall questionnaire (.584) in that country. Thus, without correcting for recall bias, the separability estimates from Tanzania using a different data source (but the same questionnaire method) replicate earlier findings.

Dillon and Barrett (2017) also compare the β_1 estimates across countries to determine the "depth of market failure". For the basic specification controlling for gender of the household head, Tanzania's estimate lies between the low of .33 for Uganda and high of .82 for Niger. Note that the magnitude of the cross-country range estimated in that paper is quite similar to the treatment effect of weekly visits estimated here for Tanzania (.44). Thus, the effect of measurement error on tests of separation explains nearly the entire gap in cross-country

estimates of the elasticity of labor with respect to household size. Put differently, given the null treatment effect in Ghana and large treatment effect in Tanzania, the sizable estimated differences between countries in tests of separation cannot be disentangled from differential sensitivity to measurement error.

To illustrate, note that the β_1 estimates from the traditional recall module for Tanzania and Ghana are nearly identical. However, the magnitude of the separation test for the weekly visit group in Tanzania is nearly double the rate of Ghana, suggesting that measurement error causes not only an overestimation of market completeness in Tanzania, but also severely overstates market completeness relative to Ghana.

.

TABLE 1 **HOUSEHOLD CHARACTERISTICS SUMMARY STATISTICS**

	GHANA			TANZANIA			
Group/ Household Characteristics	**VISIT**	**PHONE**	**TRADITIONAL RECALL**	**VISIT**	**PHONE**	**ALTERNATIVE RECALL**	**TRADITIONAL RECALL**
HOUSEHOLD SIZE	5.667*	5.339	5.203	6.58	6.793*	6.22	6.29
	(0.190)	(0.177)	(0.170)	(0.221)	(0.220)	(0.164)	(0.199)
Head Male	0.791	0.813	0.841	71.7%	82.6%	73.4%	76.4%
	(0.027)	(0.026)	(0.024)	(0.031)	(0.026)	(0.030)	(0.029)
Head Age	47.916	45.835	46.053	50.02	50.81	50.45	52.02
	(1.027)	(1.027)	(0.990)	(1.134)	(1.073)	(0.985)	(1.047)
Max Household Education:							
Primary or Less	40.7%	44.6%	46.7%	66.5%	63.4%	74.3%	67.5%
	(0.034)	(0.034)	(0.034)	(0.032)	(0.033)	(0.030)	(0.032)
More than Primary	59.3%	55.4%	53.3%	30.7%	33.8%	25.7%	32.1%
	(0.034)	(0.034)	(0.034)	(0.032)	(0.032)	(0.030)	(0.032)

Note: * indicate statistical difference in means compared to Traditional Recall group at the 10% significance level. Standard errors are reported.

TABLE 2: SUMMARY STATISTICS	GHANA			TANZANIA			
Group/ Season-wide hours	Visit	Phone	Traditional Recall	Visit	Phone	Alternative Recall	Traditional Recall
PER PERSON							
HOUSEHOLD LABOR	337.8	477.8***	333.6	200.1***	223.9***	357.4***	286.8
	(10.9)	(13.4)	(18.9)	(6.5)	(7.3)	(16.0)	(10.9)
NUMBER PERSON-PLOTS	846	722	630	900	928	651	608
PER PLOT							
HOUSEHOLD LABOR	435.9**	571.8***	374.0	182.9***	220.5***	444.1***	345.2
	(18.4)	(19.9)	(21.0)	(7.4)	(9.7)	(22.5)	(14.3)
HIRED LABOR	74.0	30.4***	85.0	9.4*	17.3	9.5	12.6
	(5.4)	(3.5)	(13.3)	(1.0)	(4.3)	(1.5)	(1.9)
TOTAL LABOR	509.9	602.2***	458.9	192.2***	237.8***	453.6***	357.8
	(20.9)	(21.1)	(25.1)	(7.7)	(12.0)	(22.5)	(14.2)
NUMBER OF PLOTS	641	610	562	985	942	524	505
PER ACRE							
HOUSEHOLD LABOR	216.4***	263.9***	134.6	335.5**	399.8	588.0*	432.5
	(11.6)	(14.0)	(11.4)	(28.7)	(39.7)	(72.3)	(36.8)
HIRED LABOR	28.3	25.9	28.7	13.2	15.1	10.5	12.9
	(2.6)	(2.046)	(4.8)	(1.9)	(4.9)	(2.1)	(2.6)
TOTAL LABOR	244.8***	289.8***	163.3	348.7**	414.9	598.5*	445.4
	(12.2)	(14.6)	(12.4)	(28.8)	(40.2)	(72.2)	(36.6)
AVERAGE ACRES	8.4	8.5	8.0	4.0***	3.9***	2.5	2.5
	(0.5)	(0.5)	(0.4)	(0.3)	(0.30)	(0.2)	(0.2)
AVERAGE PLOTS PER HOUSEHOLD	3.0***	3.0***	2.6	5.2****	4.9***	2.8	2.8
	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)
NUMBER OF HOUSEHOLDS	225	224	227	212	213	218	212
PER HOUSEHOLD							
HOUSEHOLD LABOR	1313.4***	1641.7***	925.8	843.9	959.6**	1067.4**	785.3
	(63.7)	(71.6)	(86.4)	(48.4)	(57.2)	(103.8)	(58.0)
HIRED LABOR	211.5	168.2	210.3	46.1*	68.7*	23.6	31.0
	(19.5)	(13.7)	(37.8)	(5.9)	(18.8)	(4.8)	(5.6)
TOTAL LABOR	1524.9***	1809.9***	1136.2	889.9	1028.3**	1091.0**	816.3
	(73.5)	(75.3)	(95.7)	(50.3)	(65.0)	(103.9)	(58.20)
NUMBER OF HOUSEHOLDS	225	224	227	212	213	218	212

*, **, *** denotes statistical difference between treatment groups means and traditional recall group mean using a t-test at 1, 5, 10% levels respectively.

Table 3: Ghana Ṡeparability Tests

Dep. Var: Log Household Labor Demand	[1]	[2]	[3]
log(hh size)	0.599***	0.554***	0.488***
	(0.072)	(0.115)	(0.121)
log(acre)	0.449***	0.434***	0.586***
	(0.044)	(0.039)	(0.079)
Head Gender	0.055	0.115	0.107
	(0.097)	(0.086)	(0.084)
Share of HH: Adult Female	0.842***	0.783***	0.759***
	(0.211)	(0.188)	(0.189)
Share of HH: Adult Male	0.852***	0.758***	0.732***
	(0.187)	(0.170)	(0.169)
Treated		0.647***	0.937***
		(0.187)	(0.216)
TreatedXlog(hh size)		0.002	0.083
		(0.113)	(0.120)
TreatedXlog(acre)			-0.225***
			(0.084)
$\beta_1 + \beta_7$		0.556***	0.571***
		(0.063)	(0.062)
N	676	676	676

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively. Robust standard errors are reported in the parentheses. Treated is equal to one if the household is in the visit or phone group and zero if the household is in the traditional recall group. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

Table 4: Ghana Separability Tests- By Treatment Arm

Dep. Var: Log Household Labor Demand	Visit Only	Phone Only	By Treatment
log(hh size)	0.530***	0.540***	0.556***
	(0.127)	(0.125)	(0.116)
Treated	0.537**	0.733***	
	(0.211)	(0.197)	
TreatedXlog(hh size)	-0.001	0.020	
	(0.126)	(0.118)	
log(acre)	0.474***	0.457***	0.432***
	(0.049)	(0.057)	(0.040)
Head Gender	0.148	0.165	0.107
	(0.112)	(0.126)	(0.086)
Share of HH: Adult Female	0.749***	0.668***	0.774***
	(0.258)	(0.242)	(0.187)
Share of HH: Adult Male	0.707***	0.671***	0.758***
	(0.240)	(0.217)	(0.169)
Visit			0.554***
			(0.212)
VisitXlog(hh size)			-0.013
			(0.126)
Phone			0.717***
			(0.194)
PhoneXlog(hh size)			0.032
			(0.117)
$\beta_1 + \beta_7$	0.529***	0.560***	
	(0.097)	(0.078)	
N	452	451	676

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively. Robust standard errors are reported in the parentheses. Treated is equal to one if the household is in the visit group in the visits only column and equal to one if the household is in the phone group in the phone only column, and zero if the household is in the traditional recall group. In the By Treatment column, Visit is equal to one if the household is in the visit group and zero otherwise and Phone is equal to one if the household is in the phone group and zero otherwise. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

Table 5: Tanzania Separability Tests

Dep. Var: Log Household Labor Demand	[1]	[2]	[3]
log(hh size)	0.776***	0.550***	0.560***
	(0.083)	(0.139)	(0.141)
log(acre)	0.361***	0.371***	0.341***
	(0.038)	(0.039)	(0.065)
Head Gender	0.152*	0.151*	0.151*
	(0.088)	(0.087)	(0.087)
Share of HH: Adult Female	0.527**	0.464**	0.452**
	(0.221)	(0.223)	(0.224)
Share of HH: Adult Male	0.317	0.384*	0.379
	(0.237)	(0.229)	(0.230)
Treated		-0.648**	-0.656**
		(0.271)	(0.272)
TreatedXlog(hh size)		0.349**	0.334**
		(0.146)	(0.148)
TreatedXlog(acre)			0.044
			(0.074)
$\beta_1 + \beta_7$		0.899***	0.894***
		(0.090)	(0.089)
N	574	574	574

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively. Robust standard errors are reported in the parentheses. Treated is equal to one if the household is in the visit or phone group and zero if the household is in the traditional recall group. The alternative recall group is omitted. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

Table 6: Tanzania Separability Tests- By Treatment Arm

Dep. Var: Log Household Labor Demand	Visit Only	Phone Only	Treatments
log(hh size)	0.584***	0.494***	0.580***
	(0.149)	(0.148)	(0.135)
log(acre)	0.389***	0.352***	0.370***
	(0.045)	(0.051)	(0.038)
Treated	-0.910***	-0.441	
	(0.291)	(0.325)	
TreatedXlog(hh size)	0.443***	0.282	
	(0.155)	(0.174)	
Head Gender	0.148	0.308**	0.219***
	(0.101)	(0.123)	(0.083)
Share of HH: Adult Female	0.659**	0.461	0.635***
	(0.285)	(0.285)	(0.217)
Share of HH: Adult Male	0.288	0.433*	0.635***
	(0.279)	(0.261)	(0.217)
Visit			-1.032***
			(0.306)
VisitXlog(hh size)			0.509***
			(0.161)
Phone			-0.407
			(0.328)
PhoneXlog(hh size)			0.258
			(0.174)
Alternative Control			-0.690*
			(0.352)
AltXlog(hh size)			0.435**
			(0.192)
$\beta_1 + \beta_7$	1.027***	0.776***	
	(0.119)	(0.121)	
N	385	377	770

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively. Robust standard errors are reported in the parentheses. Treated is equal to one if the household is in the visit group in the visits only column and equal to one if the household is in the phone group in the phone only column, and zero if the household is in the traditional recall group. In the By Treatment column, Visit is equal to one if the household is in the visit group and zero otherwise and Phone is equal to one if the household is in the phone group and zero otherwise. The alternative recall group is included as additional indicator variable. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

Table 7: Robustness Check – Stepwise Add Control Variables

Dep. Var: Log Household Labor Demand	Ghana				Tanzania			
	Baseline	[1]	[2]	[3]	Baseline	[1]	[2]	[3]
log(hh size)	0.488***	0.476***	0.497***	0.499***	0.560***	0.542***	0.512***	0.525***
	(0.121)	(0.122)	(0.123)	(0.120)	(0.141)	(0.142)	(0.143)	(0.145)
Treated	0.586***	0.926***	0.948***	0.974***	-0.656**	-0.688**	-0.680**	-0.670**
	(0.079)	(0.219)	(0.218)	(0.219)	(0.272)	(0.277)	(0.279)	(0.289)
TreatedXlog(hh size)	0.107	0.091	0.081	0.074	0.334**	0.349**	0.347**	0.342**
	(0.084)	(0.121)	(0.120)	(0.118)	(0.148)	(0.149)	(0.150)	(0.154)
log(acre)	0.759***	0.597***	0.604***	0.599***	0.341***	0.349***	0.346***	0.338***
	(0.189)	(0.079)	(0.080)	(0.077)	(0.065)	(0.063)	(0.063)	(0.063)
TreatedXlog(acre)	0.732***	-0.225***	-0.228***	-0.231***	0.044	0.033	0.034	0.046
	(0.169)	(0.084)	(0.084)	(0.082)	(0.074)	(0.072)	(0.071)	(0.072)
Head Gender	0.937***	0.115	0.112	0.109	0.151*	0.171**	0.176**	0.186**
	(0.216)	(0.085)	(0.084)	(0.084)	(0.087)	(0.086)	(0.085)	(0.086)
Share of HH: Adult Female	0.083	0.716***	0.742***	0.745***	0.452**	0.449**	0.408*	0.417*
	(0.120)	(0.187)	(0.190)	(0.189)	(0.224)	(0.222)	(0.225)	(0.223)
Share of HH: Adult Male	-0.225***	0.720***	0.757***	0.831***	0.379	0.332	0.276	0.273
	(0.084)	(0.169)	(0.172)	(0.170)	(0.230)	(0.230)	(0.231)	(0.229)
Fixed Effects:								
Village Level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Soil Quality & Type		Yes	Yes	Yes		Yes	Yes	Yes
Maximum HH Education			Yes	Yes			Yes	Yes
Off-Farm Income				Yes				Yes
$\beta_1 + \beta_7$	0.571***	0.567***	0.578***	0.573***	0.894***	0.891***	0.860***	0.867***
	(0.062)	(0.061)	(0.062)	(0.062)	(0.089)	(0.090)	(0.092)	(0.093)
N	676	676	676	676	574	573	573	573

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively. Robust standard errors are reported in the parentheses. Treated is equal to one if the household is in the visit or phone group and zero if the household is in the traditional recall group. The alternative recall group is omitted from the Tanzania sample. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

Table 8 : Robustness Check – Control Variables By Treatment Arm

Dep. Var: Log Household Labor Demand	Ghana			Tanzania		
	Visit Only	Phone Only	By Treatment	Visit Only	Phone Only	By Treatment
log(hh size)	0.540***	0.527***	0.565***	0.512***	0.519***	0.553***
	(0.129)	(0.124)	(0.117)	(0.152)	(0.157)	(0.140)
Treated	0.622***	0.691***		-0.903***	-0.520	
	(0.214)	(0.199)		(0.309)	(0.336)	
TreatedXlog(hh size)	-0.042	0.041		0.450***	0.312*	
	(0.127)	(0.120)		(0.157)	(0.176)	
log(acre)	0.482***	0.453***	0.438***	0.379***	0.348***	0.361***
	(0.049)	(0.056)	(0.040)	(0.046)	(0.051)	(0.039)
Head Gender	0.152	0.161	0.112	0.195*	0.341***	0.241***
	(0.112)	(0.125)	(0.086)	(0.099)	(0.119)	(0.082)
Share of HH: Adult Female	0.733***	0.607**	0.768***	0.557**	0.550*	0.611***
	(0.259)	(0.239)	(0.187)	(0.279)	(0.282)	(0.218)
Share of HH: Adult Male	0.826***	0.737***	0.855***	0.090	0.290	0.494**
	(0.249)	(0.214)	(0.170)	(0.281)	(0.266)	(0.219)
Visit			0.647***			-1.051***
			(0.214)			(0.315)
VisitXlog(hh size)			-0.056			0.512***
			(0.127)			(0.162)
Phone			0.687***			-0.416
			(0.196)			(0.341)
PhoneXlog(hh size)			0.048			0.257
			(0.119)			(0.178)
						-0.703*
Revisit						(0.361)
						0.452**
RevisitXlog(hh size)						(0.197)
				0.195*	0.341***	0.241***
Phone				(0.099)	(0.119)	(0.082)
				0.557**	0.550*	0.611***
PhoneXlog(hh size)				(0.279)	(0.282)	(0.218)
				0.090	0.290	0.494**
Alternative Control				(0.281)	(0.266)	(0.219)
						-1.051***
AltXlog(hh size)						(0.315)
						0.512***
$\beta_1 + \beta_7$	0.498***	0.568***		0.961***	0.831***	
	(0.100)	(0.079)		(0.119)	(0.126)	
N	452	451	676	384	377	769

Note: *, **, *** denotes statistical significance at 1, 5, 10% levels respectively Robust standard errors are reported in the parentheses. The following fixed effects are included: village, soil quality and health, maximum household education, and off-farm income. reated is equal to one if the household is in the visit group in the visits only column and equal to one if the household is in the phone group in the phone only column, and zero if the household is in the traditional recall group. In the By Treatment column, Visit is equal to one if the household is in the visit group and zero otherwise and Phone is equal to one if the household is in the phone group and zero otherwise. The alternative recall group is included in Tanzania as additional indicator variable. $\beta_1+\beta_7$ is the marginal effect of household size on labor demand for the treated group.

References

Abate, G.T., A. de Brauw, J. Gibson, K. Hirvonen, and A. Wolle. 2022. "Telescoping Error in Recalled Food Consumption: Evidence from a Survey Experiment in Ethiopia." *The World Bank Economic Review*, Sep, pp. lhac015.

Abate, G.T., A. de Brauw, K. Hirvonen, and A. Wolle. 2023. "Measuring consumption over the phone: Evidence from a survey experiment in urban Ethiopia." *Journal of Development Economics* 161:103026.

Abay, K.A. 2020. "Measurement errors in agricultural data and their implications on marginal returns to modern agricultural inputs." *Agricultural Economics* 51:323–341.

Abay, K.A., G. Berhane, J. Hoddinott, and K. Tafere. 2021. "Assessing Response Fatigue in Phone Surveys.", Apr, pp. .

Abay, K.A., L.E.M. Bevis, and C.B. Barrett. 2021. "Measurement Error Mechanisms Matter: Agricultural Intensification with Farmer Misperceptions and Misreporting." *American Journal of Agricultural Economics* 103:498–522.

Anderson, E., T.J. Lybbert, A. Shenoy, R. Singh, and D. Stein. 2024. "Does survey mode matter? Comparing in-person and phone agricultural surveys in India." *Journal of Development Economics* 166:103199.

Aragón, F., D. Restuccia, and J.P. Rud. 2022. "Assessing Misallocation in Agriculture: Plots versus Farms." NBER Working Paper No. w29749, National Bureau of Economic Research, Cambridge, MA, Feb.

Arthi, V., K. Beegle, J. De Weerdt, and A. Palacios-López. 2018. "Not your average job: Measuring farm labor in Tanzania." *Journal of Development Economics* 130:160–172.

Beaman, L., and A. Dillon. 2012. "Do household definitions matter in survey design? Re-

sults from a randomized survey experiment in Mali." *Journal of Development Economics* 98:124–135.

Beegle, K., C. Carletto, and K. Himelein. 2012. "Reliability of recall in agricultural data." *Journal of Development Economics* 98:34–41.

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson. 2012. "Methods of household consumption measurement through surveys: Experimental results from Tanzania." *Journal of Development Economics* 98:3–18.

Benjamin, D. 1992. "Household Composition, Labor Markets, and Labor Demand: Testing for Separation in Agricultural Household Models." *Econometrica* 60:287.

Carletto, C., A. Dillon, and A. Zezza. 2021. *Chapter 81 - Agricultural data collection to minimize measurement error and maximize coverage*, Elsevier, vol. 5 of *Handbook of Agricultural Economics*. p. 4407–4480.

Carletto, C., S. Gourlay, and P. Winters. 2015. "From Guesstimates to GPStimates: Land Area Measurement and Implications for Agricultural Analysis." *Journal of African Economies* 24:593–628.

De Weerdt, J., J. Gibson, and K. Beegle. 2020. "What Can We Learn from Experimenting with Survey Methods?" *Annual Review of Resource Economics* 12:431–447.

Dillon, B., and C.B. Barrett. 2017. "Agricultural factor markets in Sub-Saharan Africa: An updated view with formal tests for market failure." *Food Policy* 67:64–77.

Dillon, B., P. Brummund, and G. Mwabu. 2019. "Asymmetric non-separation and rural labor markets." *Journal of Development Economics* 139:78–96.

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

Gaddis, I., G. Oseni, A. Palacios-Lopez, and J. Pieters. 2021. "Measuring Farm Labor: Survey Experimental Evidence from Ghana." *The World Bank Economic Review* 35:604–634.

Gibson, J., K. Beegle, J. De Weerdt, and J. Friedman. 2015. "What does Variation in Survey Design Reveal about the Nature of Measurement Errors in Household Consumption?" *Oxford Bulletin of Economics and Statistics* 77:466–474.

Gollin, D., and C. Udry. 2021. "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture." *Journal of Political Economy* 129:1–80.

Gourlay, S., T. Kilic, and D.B. Lobell. 2019. "A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale - Productivity relationship in Uganda." *Journal of Development Economics* 141:102376.

Jeong, D., S. Aggarwal, J. Robinson, N. Kumar, A. Spearot, and D.S. Park. 2023. "Exhaustive or exhausting? Evidence on respondent fatigue in long surveys." *Journal of Development Economics* 161:102992.

Kilic, T., and T.P. Sohnesen. 2019. "Same Question But Different Answer: Experimental Evidence on Questionnaire Design's Impact on Poverty Measured by Proxies." *Review of Income and Wealth* 65:144–165.

LaFave, D., and D. Thomas. 2016. "Farms, Families, and Markets: New Evidence on Completeness of Markets in Agricultural Settings." *Econometrica* 84:1917–1960.

Lobell, D.B., G. Azzari, M. Burke, S. Gourlay, Z. Jin, T. Kilic, and S. Murray. 2020. "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis." *American Journal of Agricultural Economics* 102:202–219.

McCullough, E.B. 2017. "Labor productivity and employment gaps in Sub-Saharan Africa." *Food Policy* 67:133–152.

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

Appendix Table 1 – Dropped Observation Summary

	GHANA			Tanzania			
	Visit	Phone	Trad. Recall	Visit	Phone	Alt. Recall	Trad. Recall
Baseline Sample	240	240	239	229	227	-	-
Ending Sample	225	224	227	212	213	218	212
Households Dropped	15	16	12	17	14	-	-
Reason for Dropping:							
Did not complete endline	11	13	6	-	-	-	-
Less than 2/3 surveys complete	3	2	-	17	14	-	-
No labor reported	-	-	5	-	-	-	-
No land reported	1	1	1	-	-	-	-

Note: The reason for dropping has been implemented in the order listed. So, some households that did not complete the endline also did not complete 2/3 of the surveys. Each successive row only includes additional households dropped. In Tanzania, no baseline was given for the alternative or traditional recall groups.

Appendix Table 2 – Average Visit and Phone Survey Length - Ghana

	Weekly Visits	*Weekly Phone Calls*
Average Interview Length (minutes)	15.260	10.371
	(0.156)	(0.108)

Note: Length of survey was not available in the Tanzania data set.