Recall Bias Revisited: Measuring Farm Labor with Mixed-Mode Surveys and Multiple Imputation

Hai-Anh H. Dang
(World Bank)
hdang@worldbank.org

Calogero Carletto
(World Bank)
gcarletto@worldbank.org

Paper prepared for the IARIW-TNBS Conference on "Measuring Income, Wealth and Well-being in Africa", Arusha, Tanzania November 11-13, 2022

Concurrent Session 6A: Agriculture

Time: Saturday, November 12, 2022 [10:30 AM - 12:00 PM]

Recall Bias Revisited:
Measuring Farm Labor with Mixed-Mode Surveys and Multiple Imputation

Hai-Anh H. Dang and Calogero Carletto*

October 2022

Abstract

Smallholder farming dominates agriculture in poorer countries. Yet, traditional recall-based surveys on smallholder farming in these countries face challenges with seasonal variations, high survey costs, poor record-keeping, and technical capacity constraints resulting in significant recall bias. We offer the first study that employs a less-costly, imputation-based alternative using mixed modes of data collection to obtain estimates on smallholder farm labor. Using data from Tanzania, we find that parsimonious imputation models based on small samples of a benchmark weekly in-person survey can offer reasonably accurate estimates. Furthermore, we also show how less accurate, but also less resource-intensive, imputation-based measures using a weekly phone survey may provide a viable alternative for the more costly weekly in-person survey. If replicated in other contexts, including for other types of variables that suffer from similar recall bias, these results could open up a new and cost-effective way to collect more accurate data at scale.

Key words: farm labor, agricultural productivity, multiple imputation, missing data, survey data, Tanzania

JEL: C8, J2, O12, Q12

* Dang (hdang@worldbank.org; corresponding author) is a senior economist in the Data Production and Methods Unit, Development Data Group, World Bank and is also affiliated with GLO, IZA, Indiana University, and International School, Vietnam National University, Hanoi; Carletto (gcarletto@worldbank.org) is the manager of the Data Production and Methods Unit, Development Data Group, World Bank. We would like to thank Kathleen Beegle, Stephen Jenkins, Nayoung Lee, Diego Zardetto, and participants at the "Method and Measurements" conference organized by IPA-Northwestern University and a seminar at the University of Cincinnati for useful comments on an earlier version. We would like to thank Amparo Palacios-Lopez for her generous help with the data. We are grateful to the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through a Knowledge for Change (KCP) Research Program.

I. Introduction

The Sustainable Development Goals (SDGs), with more than 200 indicators, offer a valuable opportunity for the global community to spearhead a more aggressive data agenda. However, the SDGs also place great pressure on countries with overstretched and underfunded statistical systems. Although more than one third of the SDG indicators rely on household surveys as their primary source of data, household survey data are often either unavailable or collected with insufficient frequency. In particular, poorer countries have fewer surveys. Examining household survey data from 154 countries over three decades, a recent study suggests that a 10-percent decrease in a country's income level is associated with almost one-third fewer surveys (Dang, Jolliffe, and Carletto, 2019). Furthermore, concerns have also been raised about poor data quality particularly in poorer African countries, which could interfere with government operations and policy recommendations (Sandefur and Glassman 2015; Jerven, 2019).[1]

Despite its crucial role in the fight against poverty and food security, the agricultural sector in sub-Saharan African countries predominantly operates on a small farm basis, with smallholder agriculture accounting for as much as 80 percent of all farms (FAO, 2009). Worldwide, about 84 percent of farms are less than two hectares and average farm size has decreased in most low- and lower-middle-income countries (Lowder, Skoet, and Raney, 2016). Consequently, accurate measurement of smallholder agricultural household labor plays an indispensable part in designing well-informed policy interventions to improve their welfare. However, collecting high-quality data

[1] Another study by Serajuddin *et al.* (2015) finds that over the period 2002-2011, of the 155 countries for which the World Bank monitors poverty data using the World Development Indicators (WDI) database, almost one-fifth (i.e., 28) have only one poverty data point and as many as 29 countries do not have any poverty data points in the same period. See also Devarajan (2013) for an overview of the statistical challenges facing African countries. This data-scarce situation pertains not only to household consumption, but also to a wide range of other outcomes, such as agricultural production and labor outcomes.

on smallholder farm labor in these poorer countries is a challenging task, which is further exacerbated because of high survey costs and limited technical capacity.

Household surveys have traditionally relied overwhelmingly on respondents' self-reporting, often based on long recall methods, in order to meet multiple data needs at affordable costs. But this method of data collection can result in various types of measurement error, even over relatively short recall periods for agricultural data (Beegle, Carletto, and Himelein, 2012; Godlonton, Hernandez, and Murphy, 2018).[2] Conducting an innovative randomized survey experiment for smallholder farm labor in Tanzania, Arthi *et al.* (2018) found that, compared to a benchmark ("gold standard") measure based on weekly in-person interviews, households in traditional surveys using long recall periods reported working up to four times as many hours per person-plot. The study also found that the number of people working in agriculture was underreported in those surveys using long recall. These survey biases could result in estimates of agricultural labor productivity per hour that are too low, which concurs with the findings in previous studies (e.g., Gollin *et al.*, 2014; De Vries, Timmer, and De Vries, 2015; McCullough, 2017).

Nonetheless, while the benchmark survey provides reliable data, it is expensive to implement and impractical in many contexts. Arthi *et al.* (2018) observed that reducing the length of recall by increasing the number of visits to ten weekly visits during the agricultural season would increase the survey cost by 139 percent compared to the traditional longer recall survey mode. Worse still, aside from survey costs, it is well-known among survey practitioners that expanding the scale of

[2] A brief example can help illustrate. Farming activities are seasonal by nature and vary within the season. The traditional household survey's labor module based on the recall method – where household members are asked to report on the typical amount of time worked on specific tasks or some distant events during the past agricultural season – may not be able to capture high-quality data on these irregular, non-salient activities. Consequently, since the number of hours spent on agricultural activities such as planting often varies *across* weeks, it would be incorrect to simply ask a survey question for the typical number of hours spent on such activities *per* week. Other types of variables, particularly in the context of sub-Saharan Africa and other low-income countries, that are found to have measurement errors include land area (Carletto, Savastano, and Zezza, 2013) or crop production (Gourlay *et al.*, 2019).

diary-type face-to-face surveys in a low-income country context also presents challenges due to constraints in capacity and manpower.

More generally, the use of alternative, complementary data collection methods based on more frequent and accurate measurement is becoming more widespread in surveys.[3] Yet, the application of these alternatives in large-scale surveys remains challenging and costly. A key challenge thus emerges on finding the right balance between cost, accuracy, and capacity for data collection using these modern methods *vis-à-vis* easier and less costly, but less accurate, traditional measures based on self-reporting and long recall. Indeed, striking the right balance would have highly relevant implications for future survey design efforts, particularly in poorer countries with capacity constraints.

In this paper, we offer a new investigation into this challenge. Specifically, we study the following questions. Given Arthi *et al.*'s (2018) finding that the benchmark "gold standard" survey (based on weekly in-person visits) is more accurate but more expensive, can we use a small sample of this benchmark survey, combined with the use of imputation methods, to improve upon the accuracy of the (long) recall survey? If yes, what is the required sample size of the benchmark survey to produce an effective imputation model for labor outcomes? Furthermore, since Arthi *et al.* (2018) also suggest that the (weekly) phone survey can provide reasonably good quality estimates for some labor indicators vis-à-vis the benchmark survey, can we use a small sample of this less expensive phone-based option to impute estimates in place of the more expensive benchmark face-to-face measure? Finally, what are the tradeoffs in terms of accuracy and cost that we should consider when implementing these different types of surveys?

[3] These include, for example, deploying portable GPS devices for better measurements of land area, or the use of mobile phones to assist in diary-keeping (or short recall surveys) for labor, agricultural production, or expenditure.

Our findings suggest that rather parsimonious imputation models can offer estimates that fall within the 95-percent confidence intervals of the benchmark "gold-standard" in-person diary-based estimates. In many cases, our estimates even fall within one standard error of the benchmark estimates. We also find that a judicious combination of a smaller sample of a high-quality benchmark survey embedded in a traditional recall survey can provide reasonably good imputation-based estimates that track the benchmark estimates. Moreover, replacing the more expensive subsample based on the benchmark in-person survey with the phone-based survey offers similar encouraging estimates and provide a viable, cost-effective alternative. Similar results hold for different variants of the traditional (long) recall survey being used as the target survey.

Our study helps advance the literature in several different ways. To our knowledge, we offer the first study in the economic literature that employs multiple imputation (MI) methods to provide imputation-based estimates of smallholder farm labor.[4] Specifically, we provide imputation-based estimates of farm labor, using the same data that was analyzed by Arthi *et al.* (2018). The imputation and validation process consists of three main steps. Firstly, we build the imputation model using data from the benchmark ("gold standard") in-person diary-based survey – hereafter also referred to as the *base* survey. Secondly, we apply the estimated parameters (and their distributions) from this imputation model to the same explanatory variables in the other samples – hereafter referred to as the *target* surveys – to provide estimated indicators of farm labor. Finally, these imputation-based estimates are validated against the survey-based estimates that are directly obtained from the *base* survey data (hereafter referred to as the "true rate"). We also repeat this

[4] Imputation methods have become increasingly more common as alternative ways to address data challenges in economics, particularly with regards to poverty. The growing literature on poverty imputation is widely considered to have begun with the seminal study by Elbers, Lanjouw, and Lanjouw (2003), which imputes from a household consumption survey into a population census to obtain disaggregated estimates of poverty. See Dang *et al.* (2019) for a recent review of this literature. Obtaining high-quality data on agricultural labor is most related to SDG goal no. 8 on full and productive employment and decent work for all (https://sdgs.un.org/goals/goal8).

exercise using different subsamples as the base survey (such as the subsample of the less costly phone survey) or the target survey (such as using variants of the traditional recall survey).[5]

We also contribute to a related, multi-disciplinary literature on finding appropriate sample sizes for imputation models. Park and Dudycha (1974) offer some theoretical guidance on selecting the appropriate sample size of the *target* survey for obtaining regression-based prediction estimates. There is, however, no empirical evidence on how large the sample size in the target survey should be to produce high-quality estimates for labor indicators.[6] But more importantly for our purposes, neither does any previous (theoretical or empirical) evidence exist on sample sizes for these outcomes with the *base* survey. That is, no existing study investigated the question: what is the smallest subsample of the more expensive benchmark survey that is sufficient? Our study thus sheds light on these practical issues that comprise an integral part of using these imputation-based approaches. Finally, we discuss various cost-benefit considerations for the different modes of data collection. Taken together, our new findings in the context of Tanzania could potentially apply to other contexts, as well as to other outcome variables with data quality that are compromised by similar recall-based measurement error issues.

This paper consists of six sections. We provide a summary of the data in the next section before discussing the MI framework and the existing theory on selecting sample sizes for imputation in Section III. We present in Section IV the estimation results for agricultural working hours and the estimated sample sizes based on theoretical evidence and simulation, using both high-quality and less-than-perfect survey data in the imputation procedures (Section IV.1). We subsequently

[5] More generally, we use the terms "base survey" and "target survey" to refer, respectively, to the survey on which we build our imputation model and the survey into which we impute. We also use the term "survey-based estimates" to refer to the estimates obtained directly from the survey data, and the term "imputation-based estimates" to refer to the estimates obtained from the imputation model.

[6] For example, Anderson *et al.* (2017) and Riley *et al.* (2020) discuss different criteria that can be used to select a good sample size for prediction respectively in psychology and medical studies. Dang and Verme (2022) offer empirical evidence that sample sizes of 1,000 households or more can provide reliable estimates for poverty imputation.

provide the estimates for the number of household members working on the farm (Section IV.2) as well as the estimates when using the phone survey as the alternative benchmark (Section IV.3). We further discuss the implications for survey implementation in Section V and finally conclude in Section VI.

II. Data Description

We briefly summarize the main features of Arthi *et al.*'s (2018) survey experiment before describing the data. Arthi *et al.* (2018) conducted a survey experiment among 854 farming households in 18 enumeration areas in the Mara region of rural northern Tanzania. They collected data on labor indicators during the 2014 *masika* – the main, long rainy season in the first half of 2014. They randomly assigned households to one of the four following survey arms within each of the 18 enumeration areas.

1. **Weekly in-person visit (Arm 1)**: weekly in-person visits for the duration of the *masika* season, followed by a personal endline survey between July and September 2014.
2. **Weekly phone call (Arm 2)**: weekly phone survey for the duration of the *masika* season, followed by a personal endline survey between July and September 2014.
3. **Recall module 1 (Arm 3)**: an in-person endline survey between July and September 2014, following the same design as the Tanzania National Panel Survey household consumption survey.
4. **Recall module 2 (Arm 4)**: an in-person endline survey between July and September 2014, with a modified module design.

Several remarks on the experiment design can be useful. First, following Arthi *et al.* (2018), we consider the weekly in-person visit sample (Arm 1) as the benchmark survey. By design, this benchmark survey involves the most intense data collection effort, helping to reduce recall bias,

and its high data quality is supported by the results of qualitative focus group discussions. However, this type of survey is the most expensive to implement.[7] For the purposes of our analysis, we consider this weekly in-person visit survey to be the "gold standard" (i.e., best approximation to the true rates).

Second, all four survey arms collect data on several labor outcomes during the past agricultural season such as the number of hours, days, and weeks worked per plot, the number of household members working in farming, the number of plots worked per person.[8] For a more focused discussion, we analyze two main indicators: the total number of hours an individual worked and the number of household members that actively engaged in farming over the past agricultural season.

Table 1 provides the summary statistics for our estimation sample by the four survey arms in the first four columns, as well as the combined sample for the three other (non-benchmark) surveys in column 5. Since we restrict our sample to working adults, there are 2,748 individuals and 842 households available for analysis.[9] Similar to Arthi *et al.* (2018), compared with the benchmark weekly visit survey (column 1), we find that the two recall surveys (columns 3 and 4) provide biased estimates at the individual level, with most of the variables in the latter two surveys being statistically significantly different from those of the former. However, the individual characteristics from the weekly phone survey (column 2) are not statistically significantly different from those of the weekly in-person benchmark survey. For household characteristics, almost no difference exists among the four surveys.

[7] Arthi *et al.* (2018) offer further discussion on other features of their surveys such as within-village randomization, Hawthorne effects, self-reporting rates, and attrition issues.
[8] Recall survey 1 (Arm 3) collects data on numbers of days spent on the plot and typical hours worked per day, while Recall survey 2 (Arm 4) collects data on numbers weeks worked on the plot, number of days worked per week and hours worked per day.
[9] This sample has 12 fewer households than in Arthi *et al.* (2018), but our summary statistics for households are almost identical to theirs.

The combined sample for the three non-benchmark survey arms (i.e., Arms 2 to 4) is unsurprisingly biased *vis-à-vis* the benchmark survey but somewhat less biased than the two recall surveys (i.e., Arms 3 and 4) for the individual characteristics. For example, the proportion of individuals who are currently enrolled in school is not statistically significantly different from that in the benchmark survey. The household characteristics in the combined sample remain similar to the benchmark survey. This suggests that we can experiment with using the combined sample for obtaining imputation-based estimates. The main advantage of the combined sample is its larger sample size, which generally allows for more accurate estimates.

We show the summary statistics for the labor outcomes of interest in Table 2. The results are broadly consistent with those for Table 1: while all the three other survey arms provide biased estimates (compared to Arm 1, our benchmark survey), the phone survey (Arm 2) is relatively less biased. In particular, the number of hours worked over the past season is overreported by 10 hours for the phone survey (i.e., 56.5 hours vs. 46 hours for the benchmark survey). In contrast, the corresponding difference is eight to more than ten times larger (i.e., 80 hours and 113 hours more) respectively for the recall 1 and recall 2 surveys (Arms 3 and 4). However, the magnitude of bias is smaller for all the three other survey arms regarding the number of household members doing farm work. Furthermore, while each of the two recall surveys still provides statistically different biased estimates for this outcome, the estimates based on the phone survey and the combined sample for the two recall surveys are no longer statistically different from those of the benchmark survey.

We will impute from the benchmark survey (Arm 1) into the phone survey (Arm 2) and various combined samples of the phone survey and the two recall surveys (Arms 3 and 4) for robustness

checks (Sections IV.1 and IV.2). As an alternative, we also impute from the phone survey (Arm 2) into the two recall surveys (Arms 3 and 4) (Sections IV.3).

III. Analytical Framework

In this section, we present the Multiple Imputation (MI) method before discussing the available evidence on the sample size for imputation models.

III.1. MI Methods

There is an established literature on missing data or multiple imputation (MI) in statistics. Official agencies such as the U.S. Census Bureau routinely use imputation to fill in important missing data on various statistics for income (Census Bureau, 2016a) and labor (Census Bureau, 2016b). Early adopters of MI methods for study of economic topics include Davey, Shanahan, and Schafer (2001) for children's family experiences and psychosocial adjustment and Jenkins *et al.* (2011) for income inequality. Douidich *et al.* (2016) offer a recent application of MI methods to poverty imputation. Yet, MI methods still remain little used in economics.[10] For this reason, we provide below a discussion of MI methods based on Rubin (1988) and Little and Rubin (2020) under an econometric framework that is widely used in poverty imputation.

Let x_j be a vector of characteristics that are commonly observed between the four surveys, where j indicates one of the four survey arms in our context.[11] Subject to data availability, these characteristics can include individual-level and household-level characteristics. Individual characteristics include variables such as age, sex, education, ethnicity, religion, language, and

[10] Further discussion on the differences between poverty imputation methods in economics and MI methods is provided in Dang *et al.* (2019). Also see Dang *et al.* (2014) for an extension of poverty imputation methods in the context of synthetic panel data.

[11] More generally, j can indicate any type of relevant survey that collects household data sufficiently relevant for imputation purposes, such as labor force surveys or demographic and health surveys. To make notation less cluttered, we suppress the subscript for each household in the following equations.

occupation. Household characteristics include variables such as household size, the number of rooms in the house, the physical quality of the house (e.g., whether its roof or wall is of good quality), and the distance from the house to the nearest facilities, such as sources of water. These variables can capture the household's income levels.[12]

High-quality data on labor indicators exist in the benchmark survey but are not available in the other surveys. Thus, let survey *1* (j=1) represent the benchmark survey. The other surveys without such data are the phone survey (j=2), the combined phone and two recall surveys (j=3), or the recall 1 survey (j=4) and the recall 2 survey (j=5). Let y_1 represent the vector of outcomes of interest in survey 1. Our objective is to impute the missing (or low-quality) labor indicators in survey j, given that these labor indicators are available in survey *1 only*, and the survey characteristics x_j are available in all surveys.

We assume that the linear projection of labor indicators on household and other characteristics (x) is given by the following linear model

$$y_j = \beta_j' x_j + \varepsilon_j \qquad (1)$$

Conditional on the x_j characteristics, the error term is assumed to follow a normal distribution $\varepsilon_j | x_j \sim N(0, \sigma^2_{\varepsilon_j})$; β_j are the vector of coefficients, for j= 1,…, 5. Equation (1) thus provides a standard linear model that can be estimated using most available statistical packages.

We make the following assumption to further operationalize our estimation framework.

Assumption 1: Let x_j denote the values of the variables observed in survey j, for j= 1,…, J, and let X_j denote the corresponding measurements in the population. Then x_j are consistent measures of X_j for all j (i.e., x_j=X_j for all j).

Assumption 1 is crucial for imputation and ensures that the sampled data in each survey are representative of the target population. Different versions of this assumption are commonly

[12] Household assets or income can also be included if such data are available.

employed in previous studies on survey-to-census and survey-to-survey imputation (Elbers *et al.*, 2003; Tarozzi, 2007; Dang *et al.*, 2019). This assumption implies that, for the four surveys we have, measurements of the same characteristics *x* are identical, as they are consistent measures of the population values. While surveys of the same design (and sample frame) are more likely to be comparable and can thus satisfy Assumption 1, these surveys may not necessarily provide comparable estimates. Examples where Assumption 1 may be violated include cases where national statistical agencies change the questionnaire for the same survey over time.[13] Violation of Assumption 1 rules out the straightforward application of survey-to-survey imputation technique and would require further investigation of estimation results.

Assumption 1 can be tested when the surveys under study are implemented in the same period. The estimation results in Table 1 as discussed in Section II above suggest that the phone survey satisfies this assumption; that is, the individual and household characteristics based on this survey are not statistically significantly different from those based on the benchmark survey. Each of the two recall surveys and the combined sample, however, do not satisfy this assumption. We return to more discussion on the estimation results when relaxing this assumption in Section IV.1.

Given Assumption 1, we can replace x_1 in Equation (1) as

$$y_j^1 = \beta_1 ' x_j + \varepsilon_1 \qquad (2)$$

for $j = 2$, 3, 4, and 5. Equation (2) thus applies the model parameters β_1 and ε_1 based on the base survey 1 to the x_j characteristics in the target surveys to obtain estimates of the labor indicators y_j^1 in these surveys.

[13] The inconsistency between different rounds of the same survey or different surveys is well documented in studies using data from both poorer and richer countries. Survey design issues that compromise the comparability of poverty estimates are found in various countries such as China (Gibson, Huang, and Rozelle, 2003), Tanzania (Beegle *et al.*, 2012), and Vietnam (World Bank, 2012). See also Angrist and Krueger (1999) for a related review of comparability and other data issues with a focus on labor force surveys in the U.S.

Since the estimated parameters are obtained using a different survey from the target surveys, we can use simulation to estimate Equation (2) as follows

$$\hat{y}_j^1 = \frac{1}{S}\sum_{s=1}^{S}(\tilde{\hat{\beta}}'_{1,s}x_j + \tilde{\hat{\varepsilon}}_{1,s}) \quad (3)$$

where $\tilde{\hat{\beta}}'_{1,s}$, $\tilde{\hat{v}}_{1,s}$, and $\tilde{\hat{\varepsilon}}_{1,s}$ represent the s^{th} random draw (simulation) from their estimated distributions, for s= 1,…, S. The variance of $\hat{y}_j^1$ can be estimated as

$$V(\hat{y}_j^1) = \frac{1}{S}\sum_{s=1}^{S} V(\hat{y}_{j,s}^1|x_j) + V(\frac{1}{S}\sum_{s=1}^{S}\hat{y}_{j,s}^1|x_j) + \frac{1}{S}V(\frac{1}{S}\sum_{s=1}^{S}\hat{y}_{j,s}^1|x_j) \quad (4)$$

As an alternative to the linear regression method offered in Equation (3), we can employ a predictive mean matching (PMM) algorithm to draw $\hat{y}_j^1$ instead from the nearest matching observation in the base survey. More formally, applying the estimated parameters from Equation (2) to the base survey itself for each simulation *s*, we have

$$\hat{y}_{1,s}^1 = \tilde{\hat{\beta}}'_{1,s}x_1 + \tilde{\hat{\varepsilon}}_{1,s} \quad (5)$$

We subsequently replace $\hat{y}_{j,s}^1$ with $\hat{y}_{1,s}^1$ such that the absolute difference $|\hat{y}_{j,s}^1 - \hat{y}_{1,s}^1|$ for each individual is minimized, drawing from five nearest neighboring observations. The estimation procedures are described in more details in the Stata manual (StataCorp, 2019).

Since the PMM algorithm is non-parametric, it does not rely on the assumption of normality of the error term ε_j and offers better estimation results where such assumption does not hold (Little, 1988). This advantage may be even more relevant in our study since the various non-benchmark surveys can potentially offer biased estimates due to their small sample sizes (even where the normality assumption holds). Consequently, the PMM imputation method is our preferred estimation method and will be employed for most of the analysis. However, we also show some estimates based on Equation (3) for comparison.

III.2. Sample Size

One practically relevant question is how large the imputation sample should be to obtain accurate estimates.[14] While a large sample can provide estimates with more accuracy and generally better statistical properties, it is also more expensive and demands more logistical and technical resources to implement than a small sample. A balance should be reached between these tradeoffs to suit the specific context.

There is no existing evidence on selecting sample sizes for the base survey. But Park and Dudycha (1974) offer some theoretical guidance on selecting the appropriate sample size of the target survey for obtaining regression-based prediction estimates. In particular, we want to find the sample size *n* such that

$$P\{(\rho^2 - \rho_c^2) \leq \varepsilon\} = \gamma \tag{6}$$

where ρ^2 is the maximum (or true) multiple correlation possible for Equation (1) in the population, and ρ_c^2 is the correlation between the predicted value using Equation (1) and the original *y* variable. ρ_c^2 is usually referred to as the squared cross-validity correlation coefficient. A good sample size would ensure that the probability of obtaining an estimate within an acceptable degree of loss of precision (ε) around ρ^2 has reasonably good power (γ). Put differently, more precision (a smaller value for ε) or more test power (a larger value for γ) requires a larger sample size.

Park and Dudycha (1974) also show that the following relationship holds for ρ_c^2 and ρ^2

$$\rho_c^2 = \frac{\rho^2}{1+\frac{p-1}{F_{1,(p-1),\delta}}} \tag{7}$$

[14] Note that this challenge of finding an appropriate sample size is in the context of predicted values based on regression models, which is different from calculating sample sizes for other purposes, such as hypothesis testing. For the latter, see Cohen (1998) for a textbook treatment. Again, our focus in this paper is on selecting sample sizes for the base survey, since most household surveys that serve as the target survey typically have a good sample size.

where $F_{1,(p-1),\delta}$ has a noncentral F distribution with the noncentrality parameter δ, and p indicates the number of variables used in the regression. This implies that we have for any positive ε

$$P\{(\rho^2 - \rho_c^2) \leq \varepsilon\} = P\left\{-(p-1)^{\frac{1}{2}}\left[\left(\frac{\rho^2}{\varepsilon}\right) - 1\right]^{\frac{1}{2}} \leq t_{(p-1),\delta} \leq (p-1)^{\frac{1}{2}}\left[\left(\frac{\rho^2}{\varepsilon}\right) - 1\right]^{\frac{1}{2}}\right\} \qquad (8)$$

In other words, after we specify some (acceptable) values for ε and γ, we can obtain the value of the noncentrality parameter δ^2 for the noncentral Student's *t* distribution with *p-1* degrees of freedom that satisfies Equation (8).

Given this value for δ^2, we can derive the sample size *n* that satisfies Equation (6) as follows

$$n = \left[\delta^2 \frac{1-\rho^2}{\rho^2}\right] + p + 2 \qquad (9)$$

Equation (9) suggests that we need a larger sample size if we want more precision or more test power (as represented by δ^2), or if the true multiple correlation (ρ^2) is low. We also need a larger sample size if we employ more variables (p) in the regression.

Yet, Equation (9) offers theoretical evidence on sample sizes for a generic target survey only. The existing literature does not offer any further empirical evidence on selecting sample sizes in the target survey regarding labor indicators. Again, and more importantly for our purposes, neither does any previous (theoretical or empirical) evidence exist on sample sizes for these outcomes with the base survey. Using Arthi *et al.* (2018) data, we help address these gaps in the literature and provide empirical evidence in Section IV on the appropriate sample sizes (for both the base and the target surveys).

IV. Estimation Results

IV.1. Agricultural Working Hours and Sample Sizes

Agricultural Working Hours

We offer estimation results for the total number of hours that individuals worked during the past agricultural season in Table 3. For comparison, we show estimation results using both imputation approaches, that is, the PMM method and the linear regression method. For comparison and robustness checks, we use three models that sequentially build upon one another. Model 1 includes individual demographic variables such as age, age squared, sex, whether the individual is currently living with his (her) spouse, whether the individual's mother has deceased, and the household size. Model 2 adds to Model 1 education and employment variables such as the number of months the individual has been away since January 2014 (up to the interview time in 2014), whether the individual is currently enrolled in school, whether the individual had agriculture as their main work since January 2014, and whether the individual visited a health care provider in the past four weeks. Finally, Model 3 adds several variables to Model 2 concerning the physical characteristics of the house such as the number of rooms, the distance (in minutes) to the nearest water source, and whether the house has a good wall, a good roof, or a good floor. We use Model 3 as our main imputation model. The estimation results for the underlying linear regression model for both imputation approaches (based on Equation (1)) are shown in Appendix 1, Table 1.1.[15]

Table 3 shows in the first two rows the survey-based estimates using the phone survey (column 1), the imputation-based estimates (columns 2 to 4), and the "true rate" based on the benchmark survey in the last row (column 1). The survey-based estimates using the phone survey is 56.5 hours, which is a biased estimate of the true rate of 46 hours as discussed earlier. For the PMM method, Model 1 yields an estimated 46.4 working hours for the past agricultural season, while the corresponding figures for Models 2 and 3 are roughly 45 hours. These imputation-based

[15] Several variables related to the physical characteristics of the house in Model 3 are not statistically significant (Appendix 1, Table 1.1). But Rubin (1987) suggests that such variables can still be included in the imputation model if there is reason to believe that they can help improve imputation precision.

estimates using the PMM method all fall within the 95 percent confidence intervals (CIs) around the true rate. In fact, they even fall within the standard error of 1.6 hours around the true working hours of 46. Consistent with our earlier discussion, the linear regression method performs slightly worse than the PMM method, with its estimate for Model 3 falling outside the one standard-error bandwidth around the true rate.

Sample Sizes

We turn next to examining the question of how large the appropriate sample sizes for the imputation model should be. We start first with showing in Table 4 the estimates for the sample size of the target survey, using the existing theoretical results offered in Park and Dudycha (1974), and compare them with the estimates based on our own simulations. These combined theoretical and empirical results can provide useful comparison for the simulations for the sample size of the base survey.

Using Equations (6) and (9), we calculate the sample sizes where ε ranges from 0.01 to 0.05, and γ ranges from 0.90 to 0.99.[16] We also assume that ρ^2 is 0.18 and the number of predictors p is 15, which are the parameters obtained under Model 3 in Appendix 1, Table 1.1. Table 4 suggests that the minimum sample size is 299 observations (where ε and γ are respectively 0.05 and 0.90), and a reasonably good sample size can be just 342 observations (where ε and γ are respectively 0.05 and 0.95). On the other hand, a sample size of 2,346 can offer the best precision level and the maximal power (where ε and γ respectively equals 0.01 and 0.99). Table 4 also provides the estimated sample sizes for different combinations of precision and power levels.

[16] The values 0.05 (or smaller) and 0.90 (or larger) are usually considered good values for ε and γ respectively (DeGroot and Schervish, 2012; Pituch and Stevens, 2016).

Figure 1.1 offers the empirical evidence on the target sample size by plotting the number of working hours during the past agricultural season against the target sample size (given the base sample size of 761 individuals from the benchmark survey). Since we have only 784 observations for the phone survey, we consider a range of 100 to 750 observations for the target sample size. Figure 1.1 shows that all estimates (the green line) fall within the 95% CIs of the true rate (the gray range) and fluctuate less at a sample size of 300 or larger. Estimates appear to move closer to the true rate (the dashed red line) at larger samples. These results are consistent with the theoretical results discussed with regards to Table 4 above.

But a more relevant question for us is whether these results for the target sample size hold for the base sample size? Figure 1 offers an answer to this question by plotting the number of working hours during the past agricultural season against the base sample size (given the target sample size of 784 individuals from the phone survey). As opposed to the results for the target sample size shown in Figure 1.1, estimates fluctuate more, including falling somewhat outside the 95% CIs of the true rate, and not stabilizing until a sample size of 450 or more. However, similar to Figure 1.1, estimates grow closer to the true rate for larger samples. In summary, obtaining good imputation results appears to require a somewhat larger sample size for the base survey than for the target survey (i.e., 450 versus 300). But the results regarding choosing sample sizes for the base and target surveys are broadly and qualitatively consistent.

Employing Less-than-ideal Data for Imputation

The estimation results in the previous discussion are obtained based on the satisfaction of Assumption 1 (i.e., the individual and household characteristics of the phone survey have the same distributions as those of the benchmark survey). We now relax this assumption and examine whether estimation results still hold in several different scenarios.

First, we use as the target survey the combined sample of the phone survey and the two recall surveys (which does not satisfy Assumption 1 as Table 1 indicates; put differently, we impute from Arm 1 into Arms 2 to 4). The estimates, shown in Table 5, are still within the 95 percent CIs of the true rate. Furthermore, the estimates based on Models 2 and 3, for both the PMM and linear regression methods, even fall within one standard error of the true rate.

Second, we further disaggregate the target survey. We use as the target survey each recall survey as well as the combined sample of the two recall surveys (i.e., we impute from Arm 1 into Arm 3 or Arm 4 or both these arms). Estimation results, shown in Appendix 1, Table 1.2, still perform reasonably well. With the exception of recall survey 2, the estimates for Models 2 and 3 still fall within the 95 percent CIs of the true rate. The estimates using recall survey 1 under Models 2 and 3 also fall within one standard error of the true rate. This result can be explained by the fact that, as seen in Table 1, recall survey 2 is more biased than recall survey 1. In particular, three variables – the proportion of individuals with main work in agriculture, the proportion of individuals that visited a health care provider in the past 4 weeks, and the proportion of houses with a good floor – are strongly statistically different from the benchmark survey in recall survey 2, whereas these variables are not statistically different from the benchmark survey in recall survey 1.

Finally, Figure 2 subsequently plots the number of working hours against the base sample size for the combined target sample of the phone survey and the two recall surveys. Compared to Figure 1.1, estimates now appear to stabilize at a slightly smaller base sample size of 200 or more. This result is expected, since the target sample size is 1,987 for Figure 2 and almost three times as large as the corresponding figure of 761 for Figure 1.1. It is also consistent with the theoretical evidence in Park and Dudycha (1974) that a larger sample size for the target survey can lead to more

estimation precision (Section III.2).[17] Overall, these results suggest that our proposed imputation method remains robust even where Assumption 1 is violated to a certain degree.

IV.2. Extension to Another Labor Outcome

Table 6 considers another labor indicator, which is the number of household members that were actively involved in farm work during the past agricultural season. Given our results discussed above, we employ both types of data for imputation, the phone survey (row 1) and the less-than-ideal combined sample of the phone and the two recall surveys (row 2) for further comparison (i.e., we impute from Arm 1 into Arm 2 and Arms 2 to 4). Estimates for both types of data, under all three models, perform well and fall within one standard error of the true rate. Table 6 uses the PMM method, but estimates using the alternative linear regression show rather similar results (Appendix 1, Table 1.4). In fact, using as the target survey either the phone survey (Arm 2) or the two recall surveys (Arms 2 to 4) provides qualitatively similar results (Appendix 1, Table 1.5).

IV.3. Using the Phone Survey as the Alternative Benchmark

We turn next to examining whether we can substitute the weekly phone survey as an alternative (cost-saving) benchmark for the "gold standard" weekly in-person survey. Put differently, we pretend that the phone survey provides the "true" rates, and we can use it (as the base survey) to impute into the combined recall surveys (as the target survey). The estimation results, shown in Table 7 and Table 8 respectively for the number of agricultural hours and the number of household

[17] As an example, Table 4 shows that given the same value of γ, increasing the sample size by around 3 times helps reduce the value of ε from 0.3 to closer to 0.1. Further experiments with using various combinations of the available surveys as the target survey offer qualitatively similarly results (such as combining the benchmark and the phone surveys and combining all the four surveys respectively shown in Figures 1.2 and 1.3 in Appendix 1).

members doing farm work, are encouraging. All estimates fall within the 95 percent CIs of the true rates for both outcomes. Even better, three-fourths of the estimates even lie within one standard error of the true rates, except for the estimates using the PMM method for the number of household members doing farm work (Table 8, row 1).[18]

When we further employ each of the two recall surveys separately as the target survey, the estimation results are qualitatively similar (Tables 1.7 and 1.8, Appendix 1). The estimates are slightly better for the number of household members doing farm work for recall survey 1 (Table 1.8, Appendix 1).[19]

V. Implications for Survey Design

In the preceding section, we combine more accurate measures on a subsample of households – whether through weekly in person or phone-based interviews – with self-reported recall information – more commonly collected at scale in large household surveys – through advanced imputation techniques. Our estimation results for Tanzania provide supportive evidence for this approach and indicate that we can obtain reasonably good measures of small farm labor indicators while also reducing the cost of data collection. Below, we discuss three considerations for survey design: accuracy, cost, and capacity.

Regarding accuracy, traditional recall survey methods could yield severe survey bias, as highlighted in Arthi *et al.* (2018). Indeed, our analysis above suggests that the number of agricultural work hours can be overreported by a factor of 2.7 to 3.5 times based on traditional recall survey methods (Table 2). If we assume that work hours are accurately measured in the non-

[18] It is possible that this increased accuracy is due to higher quality data for this outcome in the phone survey. The weekly phone survey provides an estimate on the number of household members doing farm work that is not statistically significantly different from that based on the more expensive benchmark survey (Table 2).
[19] We obtain the estimates for Tables 1.7 and 1.8 (Appendix 1) using the PMM method. Using the linear regression method provides somewhat better results (available upon request).

agricultural sectors, this would result in agricultural productivity per hour being underestimated by the same figures, relative to the other sectors. These figures are consistent with earlier findings on the agricultural productivity gap in the region. In particular, Gollin *et al.* (2014) find a productivity gap of more than 3 times between the agricultural sector and the non-agricultural sector in Africa.[20] Our proposed method can help address this survey accuracy bias and produce imputation-based estimates that are close to those based on the benchmark survey (as shown in Tables 3, 5, and 6). While the quality of the phone survey can be further improved, estimates using the phone survey as an alternative benchmark are quite encouraging (Tables 7 and 8).

Our proposed method may also be cost-effective, particularly in contexts where there is a need to scale up the sample size. Reviewing the literature on survey designs, De Weerdt *et al.* (2020) observe that benchmark surveys are typically not implemented in larger-scale surveys because they would be prohibitively expensive. For a concrete example, Table 1.6 in the Appendix (based on Arthi *et al.*'s (2018) study) shows the cost increase relative to the baseline survey for the benchmark and phone survey. Conducting one additional benchmark survey results in a cost increase of 14 percent (measured in terms of the cost of the baseline survey in the cited study) but conducting one additional phone survey is less than half as expensive, at 6 percent. Similarly, conducting 10 more benchmark surveys requires an additional cost increase of 139 percent, but the corresponding figure for phone surveys only requires an additional cost increase of 54 percent. Our imputation approach suggests that we can economize on the number of benchmark in-person weekly surveys to save costs. For example, if we conduct one benchmark survey and nine phone-based interviews instead of 10 benchmark surveys, we can reduce survey costs by roughly 70

[20] Other studies offer estimates that fall in a similar range. For example, while De Vries *et al.* (2015) find the agricultural productivity gap to range from 4 to 6.5 times, compared with the services sector and the industry sector respectively for 11 sub-Saharan African countries, McCullough (2017) estimates this gap to be 2.1 times for Tanzania.

percent. Furthermore, we provide both theoretical and empirical evidence that suggests that a reasonably good sample size for the benchmark survey (for imputation) can be as low as approximately 450 observations.

In fact, for the labor outcomes where the phone survey provided comparably high quality data as those from the in-person weekly visits, our estimation results (Table 7 and Table 1.7, Appendix 1) suggest that we can cut costs even further by implementing the less expensive phone survey in combination with recall surveys. For example, if we implement 10 phone surveys instead of 10 weekly visit surveys (in combination with long recall surveys), we can reduce the total survey costs by approximately 80 percent. Notably, these savings in survey costs can increase with the number of the survey rounds. As such, in addition to concurring with Arthi *et al.*'s (2018) suggestion that the phone survey may be an attractive option for reducing error in the measurement of rural agricultural labor, we also find it to be a cost-effective option.[21] Furthermore, in contexts where the weekly in-person visits are not an option (such as in conflict situations or during the Covid-19 pandemic), the phone survey can serve as the best option available.

Our proposed method may also help address concerns with the existing constraints to data collection capacity, including timely implementing data collection, in poorer countries. As mentioned earlier, poorer countries have fewer surveys, partially as a result of their lower levels of statistical capacity (Chandy and Zhang, 2015; Cameron *et al.*, 2021). The lack of well-trained statistical staffs has long been recognized as a challenge in many poorer countries (Jerven, 2019; World Bank, 2021). As such, instead of implementing the logistically (and financially) demanding task of scaling up the traditional survey, it could be worthwhile to experiment with improving the

[21] For the main outcome variable of agricultural working hours, while estimates based on the phone survey are less accurate than those based on the weekly visits, they are still much closer to the latter than estimates based on the recall surveys (Table 2).

technical capacity of national experts to produce imputation-based estimates. Those local experts could be paired with leading experts from countries with stronger statistical capacity, with financial assistance or coordination from international development agencies, to obtain faster and more frequent estimates.[22] In order to achieve economies of scale and scope, the possibility of conducting such training at the (sub-) regional level could also be explored.

These advantages of phone surveys are clearly brought out in their recent use to cover the Covid-19 pandemic, especially in developing countries. In particular, Miguel and Mobarak (2022) suggest that economic data are not as well-regulated in poorer countries as in richer countries, so phone surveys offer a good method of tracking economic conditions during the pandemic in poorer countries. However, phone surveys may be susceptible to certain limitations such as low response rates or under-coverage. By design, phone surveys also have shorter questionnaires with much fewer variables than the typical household survey, so phone surveys may not allow full analysis as can be implemented with the regular household survey. Indeed, Jain *et al.* (2020) observe that the response rate in their phone survey was approximately 40%, which is higher than the traditional attrition rate of 20-30%. Egger *et al.* (2022) acknowledge that by design, the short duration of the phone surveys offer relatively coarse measures of income and welfare and may not capture well very poor households, who may not own phones or live in areas with low connectivity. It would thus be useful, if not a prerequisite, to clearly identify the population that a phone survey is representative of, as well as its potential limitations, before implementing imputation.

[22] This was in fact the operational model that helped build survey implementation capacity in poorer countries when they first implemented household surveys. See Deaton (1997) and Grosh and Glewwe (2000) for further discussion on the early stages of establishing household living standards measurement surveys (LSMSs) in poorer countries.

While our proposed approach clearly demands more research, these results are consistent with the burgeoning literature on employing imputation-based methods, including machine learning techniques, to combine different data sources in predicting various other welfare outcomes such as household consumption, poverty, inequality, malnutrition, consumer demand, and agricultural crop yields (Elbers *et al.*, 2003; Gibson, 2018; Bourguignon and Dang, 2019; Dang *et al.*, 2019; Lobell *et al.*, 2020; Blumberg and Thompson, 2022; McBride *et al.*, 2022). Developing the proper imputation methods and drafting a clear protocol to scale up the use of these techniques to a wide range of surveys and development outcomes could be a game-changer that facilitates the regular and more cost-effective production of indicators from large national household surveys.[23]

VI. Conclusion

We offer the first study that applies MI methods to provide affordable and more accurate imputation-based estimates of smallholder farm labor. We also investigate the appropriate sample sizes that can be employed for imputation, which has received little attention in the literature. Our findings suggest that rather parsimonious imputation models based on relative small in-person diary-based visits integrated into traditional large recall-based surveys can offer estimates that fall within the 95 percent confidence intervals of the benchmark estimates; in many cases, our estimates even fall within one standard error of the benchmark estimates. Estimation results are robust to different modelling and data assumptions.

[23] Moreover, given some additional reasonable assumptions (e.g., the same imputation model applies to another location or time period), there are other potentially useful applications of our proposed approach in other data-scarce contexts as well. For example, if high-quality data on labor outcomes cannot be collected for one geographic location (e.g., because of limited access due to conflicts), we can apply the estimated parameters from another similar location to provide imputation-based estimates. As another example, we can also impute for labor outcomes in a more recent survey round using the estimated parameters from an older survey, or vice versa to obtain a more consistent series of labor indicators.

Our findings have far-reaching implications for future survey data collection efforts. First, we demonstrate that multiple imputation is an effective tool for increasing accuracy while reducing costs of data collection in the context of smallholder farm labor, and potentially many more applications and contexts. Additionally, we show how combining different modes of data collection, e.g. by including a phone-based diary survey in a more traditional recall survey, can be a powerful approach to collect labor data in a more accurate and cost-effective manner. Overall, higher-frequency phone surveys performed quite well relative to our benchmark weekly in-person visits, holding greater promise for future data collection, as mobile coverage and digital literacy expands even in among most marginal groups and geographically remote smallholders. The cost of phone surveys will likely decrease over time with further technological advances, and possible coverage biases be reduced with increasing internet connections[24]. All in all, the findings consistently demonstrated that multiple imputation techniques combined with the parsimonious use of higher-frequency phone surveys of a sub-sample of respondents may offer the best cost-accuracy trade-offs.

Looking ahead, we anticipate that the replication of these results in other contexts and with other variables subject to recall bias could ultimately open up a new and cost-effective way to collect high-quality data at scale. More generally, monitoring the SDGs will require the well-informed use of different data sources which, if properly leveraged and made interoperable, can improve data availability while also enhancing its accuracy and cost-effectiveness.

[24] Although not an issue in the data used for this study, phone surveys may also suffer from higher non-response rates *vis à vis* face-to-face surveys; thus, attention should be paid in both minimizing non-response and correcting ex-post for possible resulting bias.

References

Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, and James R. Spletzer. (2013) "Exploring Differences in Employment between Household and Establishment Data". *Journal of Labor Economics*, 31, S129-S172.

Anderson, Samantha F., Ken Kelley, and Scott E. Maxwell. (2017). "Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty." *Psychological Science*, 28(11): 1547-1562.

Angrist, Joshua. D. and Alan B. Krueger. (1999) "Empirical Strategies in Labor Economics." In Ashenfelter, Orley and David E. Card. (Eds.). *Handbook of Labor Economics*, Vol. 3c. Amsterdam: North-Holland.

Arthi, Vellore, Kathleen Beegle, Joachim De Weerdt, and Amparo Palacios-López. (2018). "Not your average job: Measuring farm labor in Tanzania." *Journal of Development Economics*, 130: 160-172.

Beegle, Kathleen, Calogero Carletto, and Kristen Himelein. (2012). "Reliability of recall in agricultural data." *Journal of Development Economics*, 98(1): 34-41.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman, and John Gibson. (2012). "Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania". *Journal of Development Economics*, 98(1): 3-18.

Blumberg, Joey, and Gary Thompson. (2022). "Nonparametric segmentation methods: Applications of unsupervised machine learning and revealed preference." *American Journal of Agricultural Economics* 104, no. 3: 976-998.

Bourguignon, Francois and Hai-Anh Dang. (2019). "Investigating Welfare Dynamics with Repeated Cross Sections: A Copula Approach". *Paper presented at the WB-IARIW conference on 'New Approaches to Defining and Measuring Poverty in a Growing World'*, Washington, DC.

Cameron, Grant, Hai-Anh Dang, Mustafa Dinc, James Foster, and Michael Lokshin. (2021). "Measuring the Statistical Capacity of Nations". *Oxford Bulletin of Economics and Statistics*, 83(4): 870-896.

Carletto, Carletto, Sara Savastano, and Alberto Zezza. (2013). "Fact or artifact: The impact of measurement errors on the farm size–productivity relationship". *Journal of Development Economics*, 103: 254-261.

Carpenter, J. and M., Kenward (2013). *Multiple Imputation and its Application*. Chichester: John Wiley & Sons.

Census Bureau. (2016a). Survey of Income and Program Participation, Data Editing and Imputation. Accessed on the Internet on February 21, 2020 at

http://www.census.gov/programs-surveys/sipp/methodology/data-editing-and-imputation.html

---. (2016b). Current Population Survey, Imputation of Unreported Data Items. Accessed on the Internet on February 21, 2020 at https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/imputation-of-unreported-data-items.html

Chandy, Laurence and Christine Zhang. (2015). "Stuffing data gaps with dollars: What will it cost to close the data deficit in poor countries?" Brookings Institution, Global Economy and Development program, Op-ed.

Cohen, Jacob. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Erlbaum: Hillsdale, NJ.

Dang, Hai-Anh and Paolo Verme. (2022). "Estimating Poverty for Refugee Populations Can Cross-Survey Imputation Methods Substitute for Data Scarcity?" *Journal of Population Economics*. Doi: https://doi.org/10.1007/s00148-022-00909-x

Dang, Hai-Anh, Dean Jolliffe, and Calogero Carletto. (2019). "Data Gaps, Data Incomparability, and Data Imputation: A Review of Poverty Measurement Methods for Data-Scarce Environments". *Journal of Economic Surveys*, 33(3): 757-797.

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. (2014). "Using Repeated Cross-Sections to Explore Movements in and out of Poverty". *Journal of Development Economics*, 107: 112-128.

Davey, Adam, Michael J. Shanahan, and Joseph L. Schafer. (2001). "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation." *Journal of Human Resources*, 36: 500–519.

De Weerdt, Joachim, John Gibson, and Kathleen Beegle. (2020). "What can we learn from experimenting with survey methods?" *Annual Review of Resource Economics*, 12: 431-447.

DeGroot, Morris H. and Mark J. Schervish. (2012). *Probability and Statistics*, 4th edition. Boston: Pearson Education.

Devarajan, Shantayanan. (2013). "Africa's statistical tragedy." *Review of Income and Wealth*, 59: S9-S15.

de Vries, Gaaitzen, Marcel Timmer, and Klaas de Vries. (2015). "Structural Transformation in Africa: Static Gains, Dynamic Losses". *Journal of Development Studies*, 51:6, 674-688

Deaton, Angus. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. MD: The Johns Hopkins University Press.

Douidich, Mohamed, Abdeljaouad Ezzrari, Roy van der Weide, and Paolo Verme. (2016). "Estimating Quarterly Poverty Rates Using Labor Force Surveys: A Primer." *World Bank Economic Review*, 30(3): 475-500.

Egger, Dennis, Edward Miguel, Shana S. Warren, Ashish Shenoy, Elliott Collins, Dean Karlan, Doug Parkerson *et al.* (2021). "Falling living standards during the COVID-19 crisis: Quantitative evidence from nine developing countries." *Science Advances*, *7*(6), eabe0997.

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. (2003). "Micro-Level Estimation of Poverty and Inequality." *Econometrica*, 71(1): 355-364.

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.

Gibson, John. (2018). "Forest loss and economic inequality in the Solomon Islands: using small-area estimation to link environmental change to welfare outcomes." *Ecological Economics*, 148: 66-76.

Gibson, John, Jikun Huang, and Scott Rozelle. (2003). "Improving Estimates of Inequality and Poverty from Urban China's Household Income and Expenditure Survey". *Review of Income and Wealth*, 49(1): 53–68.

Godlonton, Susan, Manuel A. Hernandez, and Mike Murphy. (2018). "Anchoring bias in recall data: Evidence from Central America." *American Journal of Agricultural Economics*, 100(2): 479-501.

Gollin, Douglas, Lagakos, David, Waugh, and Michael E. (2014). "The agricultural productivity gap". *Quarterly Journal of Economics*, 129(2): 939–993.

Gourlay, Sydney, Talip Kilic, and David 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.

Grosh, Margaret and Paul Glewwe. (2000). *Designing Household Survey Questionnaires for Developing Countries*. Washington, DC: World Bank.

Jain, Ronak, Joshua Budlender, Rocco Zizzamia, and Ihsaan Bassier. (2020). "The labor market and poverty impacts of covid-19 in South Africa." *CASE Working Paper WPS/2020-14*.

Jenkins, Stephen P., Richard V. Burkhauser, Shuaizhang Feng, and Jeff Larrimore. (2011). "Measuring Inequality Using Censored Data: A Multiple-imputation Approach to Estimation and Inference." *Journal of the Royal Statistical Society: Series A*, 174(1): 63–81.

Jerven, Morten. (2019). "The Problems of Economic Data in Africa." In *Oxford Research Encyclopedia of Politics*.

Little, Roderick J. A. (1988). “Missing-data Adjustments in Large Surveys”. *Journal of Business and Economic Statistics*, 6: 287–296.

Little, Roderick J. A. and Donald B. Rubin. (2020). *Statistical Analysis with Missing Data*. 3rd Edition. New Jersey: Wiley.

Lobell, David B., George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan 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(1): 202-219.

Lowder, Sarah K., Jakob Skoet, and Terri Raney. (2016). "The number, size, and distribution of farms, smallholder farms, and family farms worldwide." *World Development*, 87: 16-29.

McBride, Linden, Christopher B. Barrett, Christopher Browne, Leiqiu Hu, Yanyan Liu, David S. Matteson, Ying Sun, and Jiaming Wen. (2022). "Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning." *Applied Economic Perspectives and Policy*, 44: 879–892.

McCullough, Ellen B. (2017). “Labor productivity and employment gaps in sub-Saharan Africa”. *Food Policy*, 67: 133–152.

Miguel, Edward, and Ahmed Mushfiq Mobarak. (2022). "The economics of the COVID-19 pandemic in poor countries." *Annual Review of Economics*, 14: 253-285.

Park, Colin N. and Arthur L. Dudycha. (1974). "A cross-validation approach to sample size determination for regression models." *Journal of the American Statistical Association*, 69(345): 214-218.

Pituch, Keenan A. and James P. Stevens. (2016). *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM’s SPSS*. Routledge: New York.

Riley, Richard D., Joie Ensor, Kym IE Snell, Frank E. Harrell, Glen P. Martin, Johannes B. Reitsma, Karel GM Moons, Gary Collins, and Maarten van Smeden. (2020). "Calculating the sample size required for developing a clinical prediction model." *BMJ* 368: m441.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: Evidence from African survey and administrative statistics. *Journal of Development Studies*, *51*(2), 116-132.

Serajuddin, Umar, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen. (2015). "Data deprivation: another deprivation to end." *World Bank Policy Research Paper No. 7252*. World Bank, Washington, DC.

StataCorp. (2019). *Stata: Release 16. Statistical Software*. College Station, TX: StataCorp LLC.

Tarozzi, Alessandro. (2007). “Calculating Comparable Statistics from Incomparable Surveys, With an Application to Poverty in India”. *Journal of Business and Economic Statistics* 25(3): 314-336.

World Bank. (2012). “Well Begun, Not Yet Done: Vietnam’s Remarkable Progress on Poverty Reduction and the Emerging Challenges”. *Vietnam Poverty Assessment Report 2012*. Hanoi: World Bank.

---. (2021). *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank.

Table 1. Summary Statistics

Variables	Benchmark Survey	Other Survey Types			
		Phone	Recall 1	Recall 2	Total
Individuals (Total= 2748)					
Age	30.2	31.11	35.38***	35.14***	33.62***
Proportion male	0.48	0.50	0.49	0.52*	0.50
Proportion living with spouse	0.44	0.48	0.55***	0.52***	0.51***
Proportion mother deceased	0.23	0.25	0.29**	0.29**	0.27**
Number of months away since January	0.20	0.24	0.41***	0.23	0.29*
Proportion in school	0.26	0.29	0.20**	0.18***	0.23
Proportion with main work as agriculture since January	0.66	0.63	0.67	0.72**	0.67
Proportion visit health care provider in the past 4 weeks	0.16	0.14	0.15	0.11***	0.13*
N	761	784	585	618	1987
Households (Total= 842)					
Household size	6.44	6.52	6.28	6.21	6.33
Number of rooms in the house	2.93	3.08	2.86	2.98	2.97
Minutes to water source	58.49	55.13	54.91	53.5	54.5
Proportion with good walls	0.47	0.48	0.40	0.44	0.44
Proportion with good roof	0.74	0.78	0.75	0.78	0.77
Proportion with good floor	0.22	0.32**	0.24	0.31**	0.29**
N	206	206	212	218	636

Note: Data from the endline survey are shown. ***p<0.01, **p<0.05, and *p<0.10 denote the statistical significance levels from the means of the benchmark survey. Data under the "Total" column are obtained from combining the phone survey and the two recall surveys.

Table 2. Agricultural Labor Reported over the Past Season

	Benchmark Survey	Other Survey Types			
		Phone	Recall 1	Recall 2	Total
Number of hours worked	45.97	56.50***	125.90***	159.25***	108.89***
N	761	784	585	618	1987
Number of household members doing farm work	4.3	4.28	2.76***	2.83***	3.26
N	186	194	211	218	623

Note: Data from the endline survey are shown. ***p<0.01, **p<0.05, and *p<0.10 denote the statistical significance levels from the means of the benchmark survey. Data under the "Total" column are obtained from combining the phone survey and the recall surveys.

Table 3. Imputation-based Estimates of Agricultural Hours for Individuals, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Predictive mean matching	56.50	**46.36***	**45.33***	**45.10***
	(2.10)	(2.2)	(2.4)	(2.2)
N	784	784	784	784
2) Linear regression	56.50	**46.34***	**44.97***	**44.24**
	(2.10)	(2.6)	(2.5)	(2.6)
N	784	784	784	784
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly visits)	45.97			
	(1.60)			
N	761			

Note: The survey-based estimates shown in the first two row are obtained directly from the Weekly phone survey; those shown at the bottom (or the true rate) are obtained from the Weekly visit survey. Estimates are obtained using with 50 iterations using the weekly phone interviews as the target survey. The true rate is obtained based on the data collected from weekly interview visits. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 4. Theoretical Sample Size for the Target Survey as a Function of the Population Parameters

Epsilon	Gamma		
	0.99	0.95	0.90
0.01	2346	1899	1684
0.02	1151	927	820
0.03	751	603	531
0.04	551	440	387
0.05	430	342	299

Note: Estimates are based on the formulae provided in Park and Dudycha (1974). We use the given parameters, the R2 value of 0.18 and the number of predictors of 15 under Model 3 from Table 1.1 in Appendix 1.

Table 5. Imputation-based Estimates of Agricultural Hours Using Less-than-ideal Data, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Predictive mean matching	108.89	**48.58**	**47.44***	**47.47***
	(2.72)	(1.5)	(1.4)	(1.4)
N	1987	1987	1987	1987
2) Linear regression	108.89	**48.51**	**47.20***	**46.76***
	(2.72)	(1.8)	(1.9)	(1.8)
N	1987	1987	1987	1987
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly visits)	45.97			
	(1.60)			
N	761			

Note: The survey-based estimates shown in the first two row are obtained directly from the combined sample of Recall surveys and Weekly phone survey; those shown at the bottom (or the true rate) are obtained from the Weekly visit survey. Estimates are obtained using with 50 iterations using the former surveys as the target survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 6. Imputation-based Estimates of Number of Household Members Working on the Farm for Individuals, Tanzania

Data Collection Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Phone	4.28	**4.40***	**4.39***	**4.36***
	(0.16)	(0.18)	(0.18)	(0.18)
N	194	194	194	194
2) Phone & Recall surveys	3.26	**4.33***	**4.34***	**4.30***
	(0.08)	(0.11)	(0.12)	(0.11)
N	623	623	623	623
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (benchmark survey)	4.30			
	(0.15)			
N	186			

Note: The survey-based estimates shown in the first two row are obtained directly from the combined phone and recall surveys; those shown at the bottom (or the true rate) are obtained from the benchmark survey. Estimates are obtained using the PMM method, with 50 iterations using the phone survey as the target survey. The true rate is obtained based on the data collected from the benchmark survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 7. Imputation-based Estimates of Number of Agricultural Hours Using the Phone Survey as the Benchmark, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Predictive mean matching	166.35	**78.31***	**77.72***	**77.27***
	(6.93)	(4.41)	(4.13)	(4.31)
N	429	429	429	429
2) Linear regression	166.35	**77.71***	**77.29***	**76.48***
	(6.93)	(5.55)	(6.02)	(6.42)
N	429	429	429	429
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly phone survey)	76.82			
	(4.33)			
N	194			

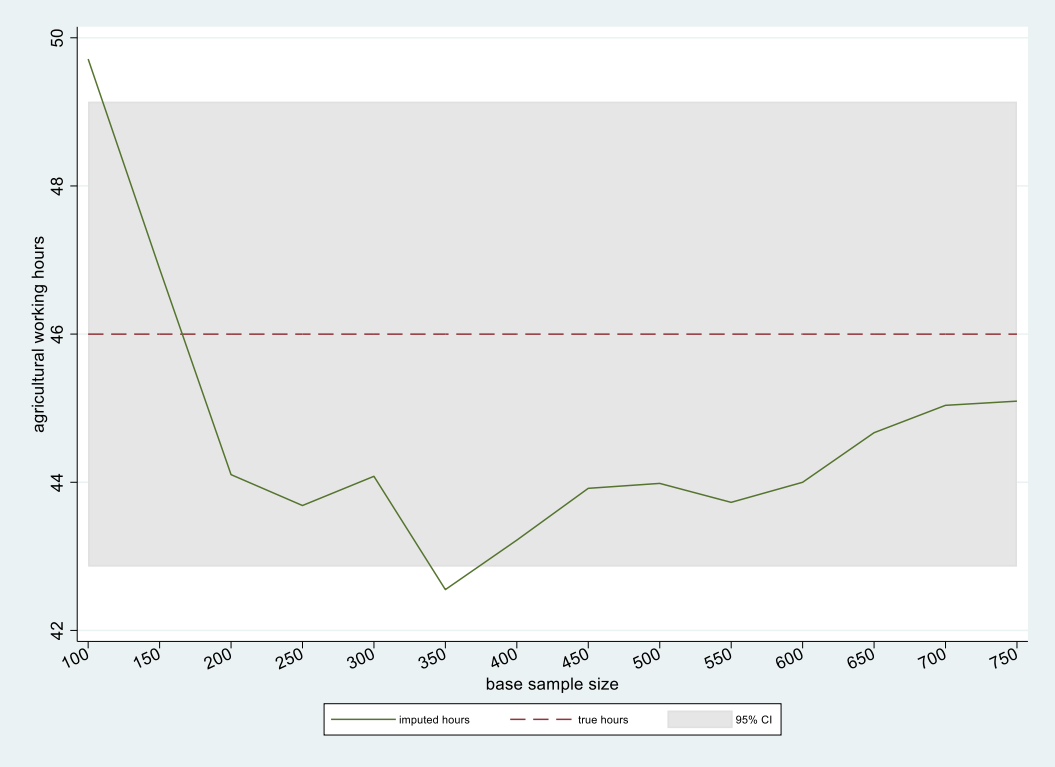
Note: The survey-based estimates shown in the first two row are obtained directly from the combined Recall surveys; those shown at the bottom (or the true rate) are obtained from the Weekly phone survey. Estimates are obtained using with 50 iterations using the former surveys as the target survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 8. Imputation-based Estimates of Number of Household Members Doing Farm Work Using the Phone Survey as the Benchmark, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Predictive mean matching	2.80	**4.11**	**4.11**	**4.09**
	(0.07)	(0.12)	(0.11)	(0.12)
N	429	429	429	429
2) Linear regression	2.80	**4.15***	**4.13***	**4.12***
	(0.07)	(0.15)	(0.16)	(0.17)
N	429	429	429	429
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly phone survey)	4.28			
	(0.16)			
N	194			

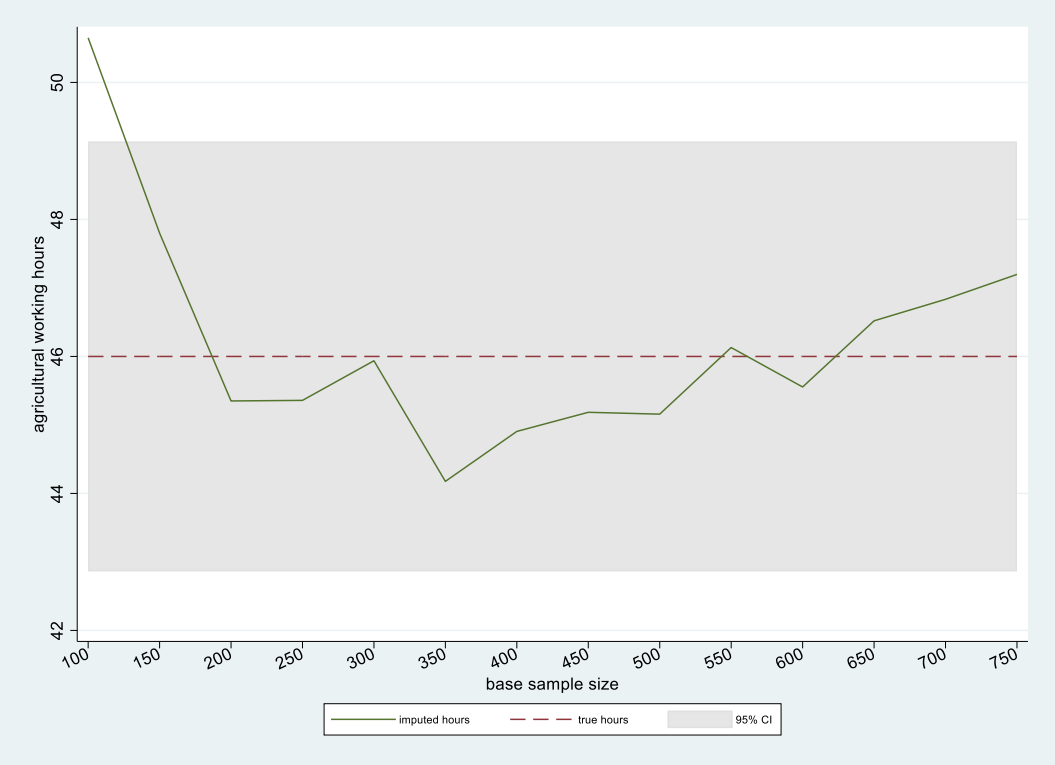
Note: The survey-based estimates shown in the first two row are obtained directly from the combined Recall surveys; those shown at the bottom (or the true rate) are obtained from the Weekly phone survey. Estimates are obtained using with 50 iterations using the former surveys as the target survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Figure 1. Imputation-based Estimates of Agricultural Hours for Different Sample Sizes of the Base (Benchmark) Survey, Tanzania



Note: The target sample size is 784 individuals from the phone survey.

Figure 2. Imputation-based Estimates of Agricultural Hours for Different Sample Sizes of the Base (Benchmark) Survey, Tanzania



Note: The target sample size is 1,987 individuals from the other three surveys.

Appendix 1. Additional Tables and Figures

Table 1.1. Estimation Model for Number of Agricultural Hours Worked over the Past Season

	Model 1	Model 2	Model 3
Age	3.026***	2.034***	2.051***
	(0.33)	(0.37)	(0.37)
Age squared	-0.031***	-0.021***	-0.021***
	(0.00)	(0.00)	(0.00)
Male	1.496	4.561*	4.688*
	(2.51)	(2.46)	(2.47)
Living with spouse	9.004***	5.270*	5.182
	(3.23)	(3.20)	(3.21)
With mother deceased	-2.326	-0.664	-0.818
	(3.55)	(3.45)	(3.46)
Household size	-0.544	-0.567*	-0.469
	(0.34)	(0.33)	(0.34)
Number of months away since January		-4.012***	-4.051***
		(1.30)	(1.31)
In school		8.342*	8.921*
		(4.95)	(5.00)
With main work as agriculture since January		33.159***	33.596***
		(4.06)	(4.09)
Visit health care provider in the past 4 weeks		-5.624*	-5.903*
		(3.42)	(3.45)
Number of rooms in the house			-1.665
			(1.17)
Minutes to water source			0.021
			(0.03)
House with good walls			2.001
			(3.06)
House with good roof			-1.376
			(3.42)
House with good floor			4.295
			(3.17)
Constant	-1.794	-5.915	-4.629
	(6.05)	(7.84)	(8.32)
Adjusted R2	0.13	0.18	0.18
N	1545	1545	1545

Note: ***p<0.01, **p<0.05, and *p<0.10. The dependent variable is the number of agricultural hours that an individual worked over the past season. The estimation sample size consists of the combined benchmark and phone surveys.

Table 1.2. Imputation-based Estimates of Agricultural Hours Different Survey Modules, Tanzania

Data Collection Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Recall 1	125.9	49.8	**47.5***	**47.4***
	(4.5)	(2.9)	(2.6)	(2.5)
N	585	585	585	585
2) Recall 2	159.2	49.8	50.3	49.2
	(6.4)	(2.6)	(2.6)	(2.7)
N	1203	1203	618	618
3) Both Recall surveys	143.0	50.2	**49.0**	**48.3**
	(4.0)	(1.9)	(1.9)	(1.8)
N	1203	1203	1203	1203
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (benchmark survey)	46.0			
	(1.6)			
N	761			

Note: Estimates are obtained using MI predictive mean matching method with 50 iterations on the specified data collection methods, including the phone survey and the two recall surveys. The true rate is obtained based on the data collected from the benchmark survey. Estimates fall within the 95% CI of the true rates are shown in bold. Standard errors are in parentheses.

Table 1.3. Theoretical Sample Size as a Function of the Population Parameters for All Survey Types

Epsilon	Gamma		
	0.99	0.95	0.90
0.01	2519	2029	1793
0.02	1204	960	843
0.03	763	601	524
0.04	539	420	363
0.05	403	309	265

Note: Estimates are based on the formulae provided in Park and Dudycha (1974). We use the given parameters, the R2 value of 0.10 and the number of predictors of 15, based on a similar model to Model 3 from Table 1.1 in Appendix 1.

Table 1.4. Imputation-based Estimates of Number of Household Members Doing Farm Work, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
Linear regression	4.28	**4.41***	**4.41***	**4.40***
	(0.16)	(0.23)	(0.23)	(0.22)
N	194	194	194	194
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (benchmark survey)	4.30			
	(0.15)			
N	186			

Note: The survey-based estimates shown in the first two row are obtained directly from the phone survey; those shown at the bottom (or the true rate) are obtained from the benchmark survey. Estimates are obtained using with 50 iterations using the phone survey as the target survey. The true rate is obtained based on the data collected from the benchmark survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 1.5. Imputation-based Estimates of Number of Household Members Doing Farm Work, Tanzania

Data Collection Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Phone	4.28	**4.40***	**4.39***	**4.36***
	(0.16)	(0.18)	(0.18)	(0.18)
N	194	194	194	194
2) Both Recall surveys	2.80	**4.32***	**4.30***	**4.30***
	(0.07)	(0.11)	(0.12)	(0.12)
N	429	429	429	429
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (benchmark survey)	4.30			
	(0.15)			
N	186			

Note: The survey-based estimates shown in the first two row are obtained directly from the phone survey and the recall surveys; those shown at the bottom (or the true rate) are obtained from the benchmark survey. Estimates are obtained using the linear regression method, with 50 iterations. The true rate is obtained based on the data collected from the benchmark survey. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 1.6. Survey Cost Increase per Household (percent)

Number of interviews	Benchmark survey	Phone survey
1	14	6
10	139	54
20	277	108
25	346	135
30	416	162

Note: This table is provided in Arthi *et al.* (2018). The costs are the cost increases in US dollars, per household, relative to the cost of an LSMS-type (baseline) survey.

Table 1.7. Imputation-based Estimates of Number of Agricultural Hours Using the Phone Survey as the Benchmark, Tanzania

Imputation Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Recall 1	144.56	**76.97***	**78.02***	**75.96***
	(7.36)	(6.12)	(6.47)	(5.94)
N	211	211	211	211
2) Recall 2	187.45	**78.16***	**78.92***	**77.38***
	(11.47)	(5.91)	(6.30)	(6.41)
N	218	218	218	218
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly phone survey)	76.82			
	(4.33)			
N	194			

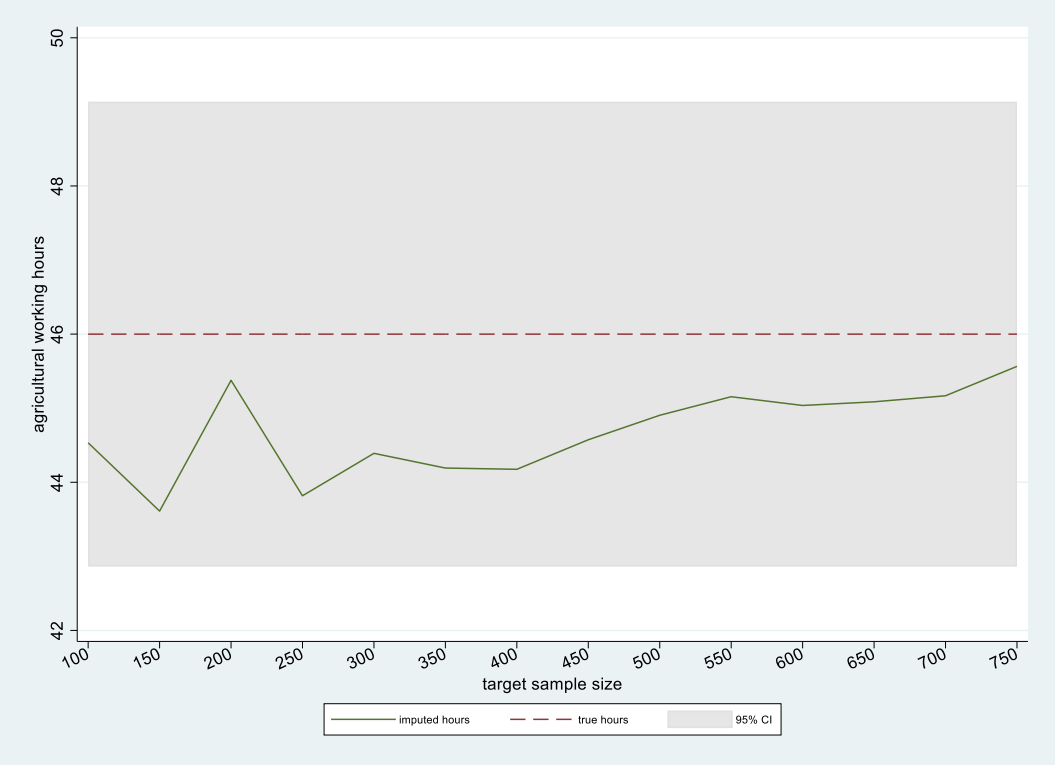
Note: The survey-based estimates shown in the first two row are obtained directly from each of the two Recall surveys; those shown at the bottom (or the true rate) are obtained from the Weekly phone survey. Estimates are obtained using with 50 iterations using MI predictive mean matching method with the weekly phone survey as the benchmark survey, and the Recall surveys as the target surveys. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Table 1.8. Imputation-based Estimates of Number of Household Members Doing Farm Work Using the Phone Survey as the Benchmark, Tanzania

Data Collection Method	Survey-based Estimates	Imputation-based Estimates		
		Model 1	Model 2	Model 3
1) Recall 1	2.76	**4.16***	**4.12***	**4.11**
	(0.11)	(0.16)	(0.17)	(0.17)
N	211	211	211	211
2) Recall 2	2.83	**4.05**	**4.08**	**4.05**
	(0.10)	(0.16)	(0.14)	(0.17)
N	218	218	218	218
Control variables				
Demographics		Y	Y	Y
Employment		N	Y	Y
House characteristics		N	N	Y
True rate (weekly phone survey)	4.28			
	(0.16)			
N	194			

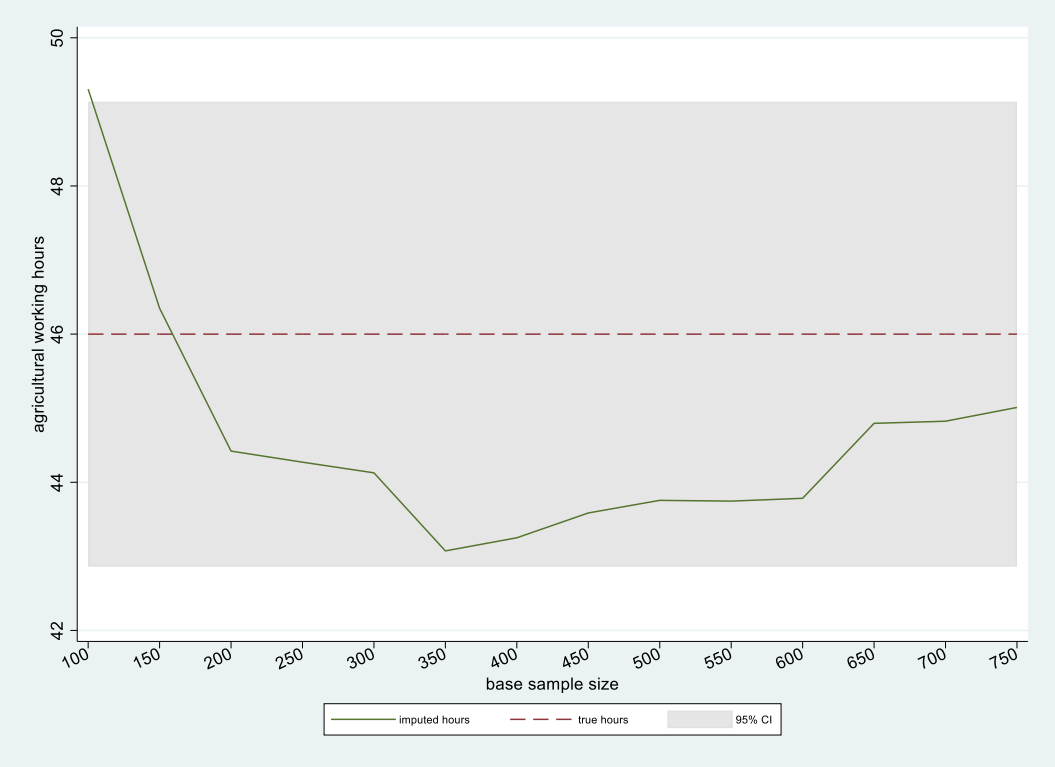
Note: The survey-based estimates shown in the first two row are obtained directly from each of the two Recall surveys; those shown at the bottom (or the true rate) are obtained from the Weekly phone survey. Estimates are obtained using with 50 iterations using MI predictive mean matching method with the weekly phone survey as the benchmark survey, and the Recall surveys as the target surveys. Estimates that fall within the 95% CI of the true rates are shown in bold; estimates that fall within one standard error of the true rates are shown in bold and with a star "*". Standard errors are in parentheses.

Figure 1.1. Imputation-based Estimates of Agricultural Hours for Different Sample Sizes of the Target (Phone) Survey, Tanzania



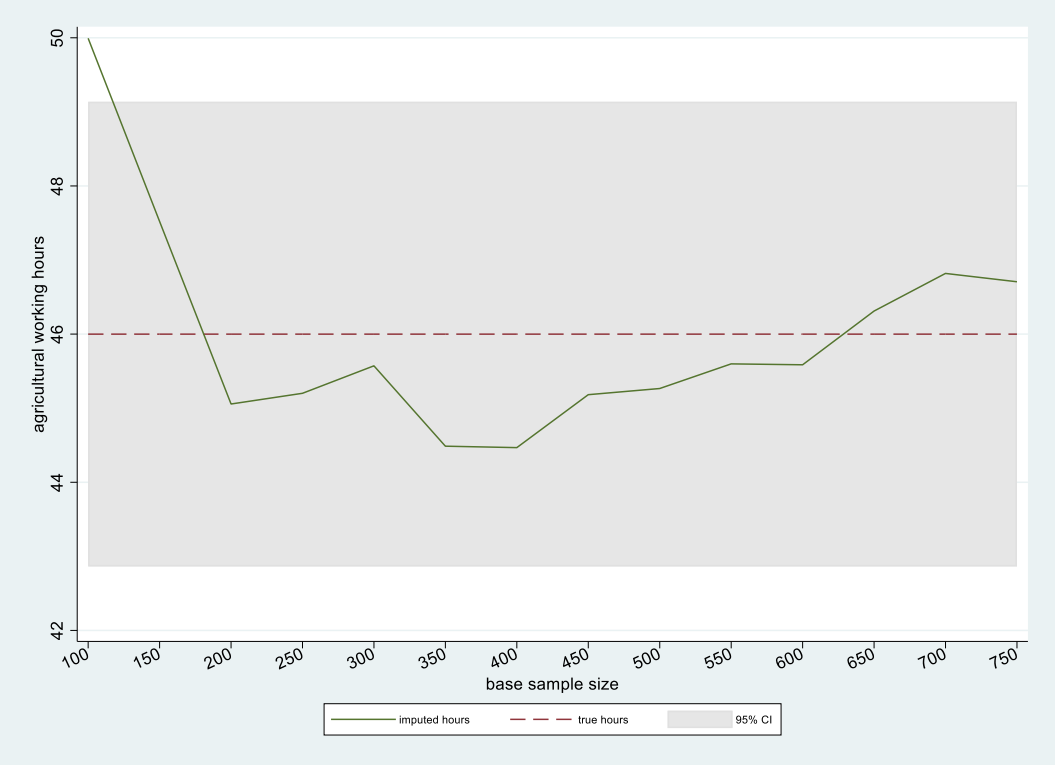
Note: The base sample size is 761 individuals from the benchmark survey.

Figure 1.2. Imputation-based Estimates of Agricultural Hours for Using the Combined Benchmark and Phone Surveys as the Target Survey, Tanzania



Note: The target sample size is 1545 individuals from combining the benchmark and the phone surveys.

Figure 1.3. Imputation-based Estimates of Agricultural Hours for Using All Four Survey Types as the Target Survey, Tanzania



Note: The target sample size is 2648 individuals from combining all the four surveys.