INTERNATIONAL CONFERENCE OF AGRICULTURAL ECONOMISTS
ICAE
29th | Milan Italy 2015
UNIVERSITÀ DEGLI STUDI DI MILANO AUGUST 8 - 14
AGRICULTURE IN AN INTERCONNECTED WORLD

Welcome To Fantasyland: Comparing Approaches To Land Area Measurement In Household Surveys

By Calogero Carletto, Sydney Gourlay*, Siobhan Murray*, and Alberto Zezza**

**World Bank*

In rural societies land is a major measure of wealth, a critical input in agricultural production, and a key variable for assessing agricultural performance and productivity. In the absence of cadastral information to refer to, measures of land plots have historically been taken with one of two approaches: traversing (very precise, but cumbersome), and farmers' self-report (cheap, but marred by large, systematic measurement error). Recently, the advent of cheap handheld GPS devices has held promise of balancing cost and precision. There are, however, concerns about how GPS measures may perform on certain types of plots, or under given measurement conditions.

Using purposely collected data from methodological validation studies conducted in Ethiopia, Nigeria, and Tanzania, this paper analyses the use of farmer self-reported area estimation against the primary objective measurement alternatives. Guided by analytical results, and with consideration for practical household survey implementation, the paper assesses the nature and magnitude of measurement error under different methods and proposes a set of recommendations for plot area measurement. Results largely point to the support of GPS measurement, with simultaneous collection of farmer self-reported areas.

IAAE

I. Introduction

Land area measurement is a fundamental component of agricultural statistics, cadastral activities, and land-related policy. Throughout history, area measurement has been used to distribute and administer lands, and collect taxes. Geometry originated in ancient Greece from the imminent need to measure land. Ancient Egyptians used the *seyat* unit in measuring land to assess peasant taxes and reallocate farmlands.[1] Land allocation in the colonization of New England began with the use of the Gunter's chain, a measure that originally varied with land quality but was eventually standardized in the 16th century allowing for consistent and comparable land area measurement.[2]

With over 70% of the developing world's poor residing in rural areas where agriculture is the primary means of livelihood, high quality agricultural data and analysis are paramount to informing policy aimed at poverty reduction (IFAD, 2010). Land is a key measure of absolute and relative farmer wealth, a critical input in production, and a key variable for normalizing agricultural input use and output measures. Although easily overlooked by analysts, the quality of land area measurement can have non-trivial implications for agricultural statistics, economics, and policy analysis. Existing literature suggests that the means of measuring agricultural land area can result in different productivity estimates as well as interfere with our understanding of agricultural relationships. For example, Carletto et al (2013b) analyze the relationship between land size and agricultural productivity with both farmer self-reported estimates and GPS measurements of land area, showing how relying on GPS area estimates results in a significantly stronger inverse productivity relationship. In a similar cross-country study, however, Carletto et al (2013a) find conflicting results, with three of four countries exhibiting a weaker (although still present) inverse productivity relationship when using GPS land area measures. The inconsistency in results adds further doubt to the use of subjective farmer self-reported areas as the impact on agricultural analysis using such estimates is potentially ambiguous.

Area measurement holds significant value in the developed country context as well. Frequent land measurement is encouraged by several developed country governments (e.g. USA, UK, Germany, Australia) at the individual farmer level in promotion of precision agriculture, whereby the production process is tailored to farmland size estimates obtained via GPS/GNSS and remote sensing. By knowing

[1] http://www.reshafim.org.il/ad/egypt/people/counting_and_measuring.htm
[2] http://www.iaao.org/uploads/norejko.pdf

exact area measurements, farmers are able to adjust input use accordingly and, thus, optimize yields while cutting down on costs. Effective land area measurement, therefore, also contributes to a more focused application of fertilizers and pesticides which could, in turn, alleviate environmental degradation and pollution. The EU FIELDFACT Project[3] is one of many schemes that aim to raise farmer awareness on the use of GPS in land measurement for the purposes of precision agriculture and more accurate and transparent subsidy claims through the EU's Common Agricultural Program which accounts for nearly half of the EU's budget. Consequently, precision agriculture (and, hence, land measurement through tools such as GPS) goes hand-in-hand with the governmental goal of boosting agricultural output while ensuring environmental sustainability and protection.

The implications of land area measurement extend well beyond agricultural productivity. Land area measurements can also be used to estimate the degree of land inequality. Unequal distribution of land has been linked to less pro-poor growth, participation in and occurrence of civil strife, and delayed long-run human capital development (Deininger and Squire, 1998; Macours 2011; André and Platteau, 1998; Baten and Juif, 2013). Carletto et al. (2013a) find that the area measurement methodology used in calculating the Gini coefficient has consequences on the level of inequality observed, with self-reported area estimates resulting in underestimated land inequality. Failure to adequately measure land limits the ability to analyze the agricultural economy and its impact on land inequality.

Land registration and titling programs require high quality land area measurement for fair program implementation. Such programs are frequently met with opposition and accusations of corruption or favoritism. Ongoing land certification reform efforts in Ethiopia have recently moved from the first stage of certification, which consisted of identification of plots by land markings and neighbor recall, to the second stage certification in which GPS measurements would replace the first stage data. A recent study analyzing the demand for the second stage certification concluded that the majority of the demand for such area measurement comes from land administrators, as households exhibited a low and declining willingness to pay over the period 2007 to 2012 (Bezu and Holden, 2013). International organizations have also emphasized the importance of objective area measurement in land registration and redistribution. In its support to the Zimbabwe Ministry of Lands and Rural Resettlement 2014-2016 Action Plan[4], the United Nations Development Programme explicitly mentions GPS and remote sensing as requirements for the successful tracking of the redistributed, post-2000 Land Reform Program parcels through the creation of a national land information database.

[3] http://www.gsa.europa.eu/introduction-and-promotion-gnss-agriculture
[4] http://www.zw.undp.org/content/dam/zimbabwe/docs/Poverty%20Reduction/UNDP_ZW_PR_Project%20Document-2014-2016%20Action%20Plan%20-%20Support%20To%20The%20MLRR.pdf

The methodological menu for collecting land area measurements is diverse and selection of the appropriate method depends on several factors. Achieving accurate area measures in a household survey setting with limited financial resources and a potentially strict fieldwork schedule is a challenging endeavor. Numerous considerations need to be balanced in designing the survey and selecting the appropriate method including the logistics of transportation to plots, the length of the full questionnaire and the likelihood of respondent and/or enumerator fatigue, and the security of teams and equipment. This paper aims to provide some elements to inform the selection of measurement methods, based on robust empirical evidence.

Since early 2013, methodological fieldwork aimed at understanding the relationship between farmer estimates, GPS measurement and the traditional compass and rope method has been conducted by the Living Standard Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) team of the World Bank in Ethiopia, Tanzania, and Nigeria. The analysis of this data along with extensive review of existing literature will be used to support specific guidelines for best practices in area measurement.

II. Methods for land area measurement in surveys

Household surveys, particularly in developing countries, have historically relied on subjective measurements of land area, and for good reason. The marginal cost of adding one or two questions to a survey that is already being administered to a household is trivial and the exercise can be completed in a matter of minutes. But, as the following sections will illustrate, there are significant trade-offs in data quality for this convenience.

Traversing, also known as the compass and rope method, is widely used in farm surveys and is often considered to be the gold standard (FAO, 1982). When properly implemented, traversing returns highly accurate measures which can also provide a benchmark against which to judge the precision of other methods. However, its implementation is time-consuming and burdensome, and is often unfeasible in the context of national household surveys and censuses. The use of Global Navigation Satellite Systems, such as the Global Positioning System, collectively referred to hereafter as GPS, can on average require as little as 28% of the time needed for compass and rope (Keita and Carfagna, 2009; Schoning et al., 2005). Keita and Carfagna (2009) find that on small plots compass and rope (“CR”) can take up to 17 times as long as GPS.

While the use of remote-sensing images and geographic information system software has the potential to be used in the household survey context, currently there are several limitations to the use of the option, particularly in the developing country environment. The spatial and temporal extent of national household surveys often makes the acquisition and processing of such high resolution imagery largely prohibitive, although with continuing advances in the technology there is potential for use of this method in the future.

Each area measurement option has unique costs and benefits that need to be carefully assessed in view of the scale of the data collection of which they are a part, the use of the data, and the characteristics of the plots to be measured and the respondents to the survey. Different methods also present different challenges in terms of their implementation: a potentially very precise method can become highly imprecise if poorly implemented in the field, or it may simply not be feasible on a larger scale. The specific limitations, challenges, and benefits of each of the abovementioned measurement methods are addressed in detail below.

A. Self-Reported Area Estimation

The quickest, most cost-effective means of assessing plot area is simply asking the farmer, "What is the area of [PLOT NAME]?". The added cost of including area estimations in an existing household survey is negligible. There are no transportation barriers, plot location and accessibility do not matter, there is no need to use both the farmer's and enumerator's time to visit the plot, and no concerns of plots being in unsafe areas. As a result, item non-response for this item is usually negligible in existing surveys. The minimal financial investment required by this method, however, does not come without challenges, both in terms of implementation and data quality.

Several factors influence the accuracy of subjective farmer self-reported estimates of area, including respondent characteristics, plot characteristics, and the land registration or titling system. Land in developing countries is often passed down from generation to generation or distributed by the community, and rarely are there property rights or documentation to inform farmers of their true land area. In many cases measurement units are not standardized and determining accurate conversion factors is a time-consuming and often ill-fated exercise. There is also evidence of enumerator effects and/or respondent learning effects which can affect data quality. The most worrying aspect of some of the measurement error associated with self-reporting is that it may be systematic, and associated with key variables of interest.

First, the precision of subjective estimates may be sensitive to respondent characteristics. A priori, it is easy to see how more educated farmers may do a better job at assessing their own land area, while absentee landlords, or respondents for which farming is only a secondary activity may be less aware of the characteristics of their plots. Using LSMS-ISA data from Uganda, Carletto et al. (2013b) analyze the determinants of the difference between farmer self-reported plot area and GPS measurements. The age of the household head has a significant impact on the accuracy of the area estimate. The perceived use of data – e.g. taxation, may induce some respondents to intentionally mis-report. Anecdotal evidence from Ethiopia purports that single-headed male households refused to allow objective measurements of their plots but offered SR estimates.

The quality of data collected through farmer self-reporting is also significantly degraded by the natural inclination of respondents to round off numbers. Distributional analysis of GPS and self-reported areas of the 2010/2011 Malawi Integrated Household Survey shows clear evidence of heaping at whole numbers and common fractions, such as 0.5 acres (see Figure 1). Carletto et al. (2013a and 2013b) find rounding to be a significant factor in the discrepancy between GPS area and farmer self-reported estimates. Plot characteristics, such as boundary delineation and the existence of property rights, are known to significantly influence the measurement bias (Carletto et al. 2013a, 2013b). Plot characteristics, such as slope or crop-type may also play a role in the ability of a farmer to estimate plot size. If a respondent purchased the plot, s/he may have documentation with the area (and an unknown method of area measurement). De Groote and Traoré (2005) assessed the accuracy of a method in which farmer self-report is elicited during a visit to the plot, and a discussion with a trained enumerator. Comparing this method to rope and compass in southern Mali they find that on average plots areas were underestimated by 11%. The observational error was strongly related to plot size, with smaller plots being overestimated and larger plots underestimated. The observational error also varied with the crop planted, being smaller for cotton fields than for cereals. The analysis was repeated at the farm level, where the bias in estimating the total area per farm was 8% underestimation.

Subjective estimates can also change significantly over time and/or based on enumerator interactions. In panel surveys one may observe improvements in the subjective measurements over time if, for example, a GPS measurement is also conducted as part of the survey and the respondents are informed of the GPS-measured area.

Farmer self-reported area estimates are influenced not only by plot and respondent characteristics, but also by a variety of cultural considerations and logistics of survey implementation. Not least of the factors complicating these estimates is the prevalence of traditional or non-standard units. When implementing a household survey, respondents are often not familiar with standard measurement units such as acres, square meters, or hectares. Depending on the country context, one is more likely to encounter a variety of traditional units, and often those units vary in size by region, or even across villages or farms. In Ethiopia, for example, one of the most common non-standards units is the *timad*, traditionally defined as the amount of land a pair of oxen can plough in one day. This measure will vary significantly by region, even by farmer. The soil texture and moisture content (clay vs sand) and plot slope will have an impact on the ease with which the oxen move. Not only that, a farmer with strong, healthy oxen will likely have a larger *timad* than a farmer with weak oxen! For Eastern Ghana, Goldstein and Udry (1999) report a correlation between self-reported and GPS measured plot size of just 0.15, which they attribute to the agricultural history of the region where local field measurements are traditionally based on length rather than area, and respondents are not accustomed to converting them to two dimensional area measures.

The existence and potentially extreme variation in traditional or non-standard units should not alone encourage the forced use of standard units, despite that fact that this will require the collection of conversion factors where they do not already exist (or were created poorly). In countries where non-standard units are most commonly used, forcing respondents to estimate plot area in a unit unfamiliar to them may result in increased measurement error.

The combination of factors influencing farmer self-reported estimates leads to large errors in area estimates. Depending on the selected respondent, that person's characteristics, the dimensions and features of the plot, and the political factors at play, the area estimate may be biased up or down at varying magnitudes. Existing literature finds, in fact, that this measurement bias is systematic. The inclusion of both farmer self-reported estimation and GPS measurement in LSMS-ISA national panel surveys allows for analysis of the accuracy of self-reported (SR) areas. GPS area is assumed to be the truer of the two measurements, and empirical evidence presented in Section III will support this assumption. Carletto et al (2013a) use LSMS-ISA data from Malawi (2010/11), Tanzania (2010/11), Niger (2011) and Uganda (2009/10) to analyze the implications of using farmer self-reported areas. A trend in the magnitude of measurement bias, defined as self-reported minus GPS area, is clearly evident across all countries. The smallest of plots (less than 0.5 acres) are systematically over-reported. The degree to which these are over-reported varies but in all countries presented the mean SR area is overestimated by at least 90% of the mean GPS area of plots in that particular area group. With increasing plot size the degree of over-estimation decreases and, eventually, converts to under-estimation for the largest plots. Carletto et al (2013b) find the same trend holds in 2005/6 data from Uganda.

The average plot sizes in the abovementioned studies are relatively small compared to what may be found in different country contexts. In a methodological study conducted in Southern Mali comparing area estimates with compass and rope measurement on larger plots (average 0.816ha, maximum 8.78ha) De Groote and Traore (2005) find that the same is true: farmers (while aided by expert observers) are inclined to over-estimate the area of plots less than one hectare, while the degree of area under-estimation increases with plot size.

Household surveys, particularly in developing countries, have historically relied on subjective measurements of land area, and for good reason. The marginal cost of adding one or two questions to a survey that is already being administered to a household is trivial and the exercise can be completed in a matter of minutes. But as technology provides alternative solutions that are both affordable and reliable, subjective measurements are increasingly being accompanied by other measures, mostly based on GPS.

B. Measurement with hand-held GPS devices

GPS measurement requires that the enumerator first traverse and clear the plot boundary with the farmer. After clearing the boundary so that it is clearly visible and can be paced easily, the enumerator then begins at a designated corner of the plot, starts the GPS area measurement function, paces the perimeter at the recommended speed (pausing at all corners to allow for coordinate capture) and completes the area measurement upon returning to the initial corner (instructions may vary by GPS unit).

The propagation of GPS technology and GPS-enabled devices offers a practical approach to objective area measurement. Kelly et al. (2008) have highlighted the use of GPS as having the potential to enable area measurement to become a much less time-intensive and costly exercise. Keita and Carfagna (2009) provide a discussion of the area measurement performance of different GPS devices compared to traversing. Their discussion is informed by a field experiment, the results of which indicate that the GPS-based area measurement is a reliable alternative to traversing and that 80 percent of the sample plots were measured with negligible error.

Advancements in GPS technology show promise for increased accuracy in the coming years. In 2011 the Russian Global Navigation Satellite System (GLONASS), which works seamlessly with the United States' GPS network, became globally operational with 24 satellites. Newer handheld GPS devices, such as the Garmin eTrex 30, are compatible with this network thereby increasing the reach, accuracy, and acquisition speed of these units. The use of differential GPS offers a tool to improve position accuracy, either in the field or through post-processing. And finally, the Wide Area Augmentation System (WAAS), a real-time correction based on ground stations, has been proven to increase position accuracy by as much as five times according to a leading manufacturer.[5] The WAAS system is only operational in North America while Europe and Asia have their own regional solutions (Euro Geostationary Navigation

[5] http://www8.garmin.com/aboutGPS/waas.html

Overlay Service (EGNOS) and Japanese Multi-Functional Satellite Augmentation System (MSAS), respectively). India's regional augmentation system (GAGAN) was cleared for navigational use in early 2014. The reach of such augmentation systems does not currently extend across Africa to any useful degree. Therefore, surveys undertaken in this region of the world cannot, at this time, take advantage of the position accuracy improvements.

Despite the great potential of GPS technology, GPS-based coordinates are subject to known types of measurement error stemming from satellite position, signal propagation, and receivers. Approximate contributions of these factors to the overall position error are significant, ranging from 0.5 to 4 meters (Hofmann-Wellenhof et al., 2008). The quality of the GPS device used also has non-negligible impact on the magnitude and distribution of measurement error (Palmegiani, 2009).

As for self-reported estimates, there is some concern that errors in GPS measures may vary systematically with some key plot characteristics, namely plot size, slope, shape, and the presence of trees. The errors in parcel boundary coordinates that underlie GPS-based area estimates have nontrivial implications. This view is supported by empirical evidence which shows that the GPS-based area measurement appears to be less precise for smaller plots, and that on average, it tends to underestimate true plot area measured by traversing (Magezi-Apuuli et al., 2005). Keita and Carfagna (2010) suggest that plots smaller than 0.5 hectares are not measured with sufficient accuracy using a handheld GPS device – a significant claim for household survey design as many rural households own or cultivate plots below this threshold. In Malawi, for example, 75% of the plots owned or cultivated by LSMS-ISA respondent households were smaller than 0.5 hectares.[6] For very large plots, on the other hand, challenges with GPS may be linked to the time and distance that needs to be walked by respondents and enumerators, and the burden that poses on the overall fieldwork (and potentially affecting the respondent's willingness to respond to other parts of the questionnaire).

Recent research by FAO points out possible effects of slope on the accuracy of GPS-based area measurement (Keita and Carfagna, 2010). Slope-related effects on area measurement are rooted in the fact that the actual area should be the horizontal projection of the plot, as opposed to the plot area itself (Muwanga-Zake, 1985). The difference between actual area and projection appear to be particularly important for slopes greater than 10 degrees (Fermont and Benson, 2011).

---

[6] Seventy-five percent of the plots owned or cultivated in the rainy season that had GPS measurements. Ninety-six percent of all plots owned or cultivated were measured with GPS. LSMS-ISA/Third Integrated Household Survey, 2010/11, Malawi. Data available online, http://go.worldbank.org/6A7GUDQ1Q0.

Bogaert, Delincè and Kay (2005), based on modeling and simulations conclude that "for GPS/EGNOS measurements made by an operator moving along the border of a field, area measurement error is linked both to the operator speed and to the acquisition rate of the GPS device. For typical field sizes found in the European Union, ranging from 0.5 ha to 5 ha, the coefficient of variation (CV) for area measurement errors is about 1% to 5%. These results depend on the field area, but they can be considered to be insensitive with respect to the field shape. They also show that field area measurement errors can be limited if an appropriate combination of operator speed and GPS acquisition rate is selected." Although, this optimal operator speed is dependent on plot size and not a single preferred pace.

Additional factors that are expected to affect the quality of GPS measures are related to the quality of signal acquisition by the receiver. These factors include the presence of dense tree canopy or cloud cover, which may interfere with the signal, and the number of satellites the receiver had acquired at the time of measurement. The higher the number of satellites, the greater the position accuracy achieved.

These concerns notwithstanding, GPS is expected to return much more precise measurement than self-reported estimates. More importantly, while the quality of GPS measures is expected to vary somewhat with some of the plot characteristics, it is not expected to vary with the characteristics of the respondents, something that we have seen plagues self-reported estimates and is an important source of concern for the analyst. One concern with GPS measures in large scale surveys, which does not apply to self-reported measures, is the rate of missingness in the data. Missingness rates of 20-30 percent are not uncommon in existing datasets, and the pattern of missingness is not random but tends to be correlated with both plot and respondent characteristics.

Kilic et al. (2013) show, with national data for Uganda and Tanzania, how plot distance from the interviewed household is the main factor determining what plots get measured, as field protocols normally include a provision not to measure plots beyond a given distance. Partly due to that, the plots that are not measured also differ systematically from those for which a GPS measure is taken in a number of (self-reported) characteristics, such as self-reported plot size, level of input use, and titling. Furthermore, some respondent characteristics are associated with higher missingness rates: plots belonging to older, less educated, poorer household heads, owning fewer plots are more likely to be measured than other plots. This is a drawback of GPS data that introduces concerns about possible biases introduced by relying on observed GPS plot measures alone.

C. The 'Gold-Standard': Compass and Rope Measurement

The compass and rope method is commonly considered to be the gold standard in objective area measurement (FAO, 1982). It does not rely on advanced technology, only basic geometry and often readily available equipment. With a compass, measuring tape, ranging poles, two to three persons and a programmable calculator or other computational tool, the area of a plot can be measured significantly more accurately than by subjective estimates. When carefully implemented, this method provides precise estimates, and it can therefore be considered as a benchmark against which to assess the precision of other methods.

The compass and rope method is burdensome and time intensive. As in the GPS measurement, this method requires that the enumerator and respondent travel to the plot, and clear its boundaries from obstacles to the extent possible. Before the measurement can begin the farmer must pace the perimeter of the plot with the enumerator in tow to ensure the measurement captures the proper area. In some cases the plot will be clearly delineated with a fence or footpath. In other quite common cases, the plot boundary will be overrun with plants, marked only by the sparse presence of a specific fencing crop or known only by the farmer. The enumerator will note the corners of the plot, where clearly available. When plots are irregularly shaped enumerators use their best judgment in declaring the corner points. The boundary of bushy plots will need to be cleared (with the permission of the farmer) prior to commencing measurement in order for the ranging poles to be visible. Only then, after travel and boundary delineation, can the enumerator start the task of measuring the plot. Moving from corner to corner, taking the compass bearings and distance between points, the enumerator can spend several hours on a single plot, depending on the size, shape, and terrain. A study by Schoning et al (2005) found the average time required for compass and rope measurement to exceed 3 hours.

To some degree, the measurement error associated with traversing is observable. The closure (or closing) error, a measure of the gap between the reported start and end points of the constructed polygon, gives an indication of the precision of the measurement. If the closing error is calculated while in the field, the measurement can be conducted again when found to be above a pre-determined threshold. The closing error will not confirm that the plot corners have been accurately assessed, however, only that the bearings and distances recorded form a full closed figure. The precision of the measurement is still subject to human error as identifying the plot corners can be a burdensome and noisy task on its own, particularly for irregularly shaped plots.

This method, despite its historical status as the benchmark measure, is not without potential error. Human error is still unavoidable as some enumerators will be more precise, have better vision, be more or less inclined to cut corners, etc. This room for human error should not be overlooked or underestimated.

Fieldwork teams will vary in capacity, work ethic, and health – all of which will impact the precision of the compass and rope measurement (as well as other forms of enumerator-executed measurement). Plots can range from clean-cut square shapes on flat plains to intricate, bushy, barely-defined plots on a steep incline. Only the most motivated teams will push themselves to mark out each corner, especially when it is 95 degrees in the shade.

Compass and rope requires intense training and demands specific qualities of the enumerators. Compass and rope is at its best when implemented by professional land surveyors, and can be problematic when working with enumerators not used to the method. Enumerator health is something that is often taken for granted but can seriously affect fieldwork, in cases for instance when impaired vision makes it difficult for them to clearly read the compass. The compass and rope method is physically demanding and, therefore, some enumerators will be better equipped for successfully and thoroughly completing the measurement. Some of these features are shared with GPS measurement, which is, however, easier to implement (even though both require adequate attention be given to training).

Neither GPS nor compass and rope measurement is void of human error. The latter, however, leaves arguably more room for intervention. While with GPS measurement an enumerator can pace the perimeter of the plot, no matter the shape, in the compass and rope technique the enumerator must first identify all corners of the plot, and when corners are not clearly defined (as is the case in many irregularly shaped plots) they must plot the "best" corner they can while trying to preserve the true area of the plot. Then at each corner the enumerator takes the compass bearings. With each additional corner, therefore, there is additional room for error in the misreading of the compass or the measurement of the distance between two corners. Misreading of the compass by one or two degrees on one corner is not likely to result in material changes to the area measurement. However, aggregated over several plot corners, these small deviations can add to a significant closing error, implying that the area calculation is not for the true plot boundaries. While field protocols usually include closing error thresholds beyond which the measurement needs to be retaken, this will add considerably to the time necessary to take the measurement.

Time can often be the scarcest resource during survey implementation. Not only does a longer fieldwork period coincide with a higher price tag but in some studies seasonality and the timing of each visit are critical. Depending on the sample size, compass and rope measurement may be prohibitively time-intensive. In addition to the cost considerations, enumerator fatigue may quickly set in the more time- and labor-intensive the in-field measurement. Existing studies comparing the time use for GPS and compass and rope find that compass and rope can take approximately 3.5 times as long as required for GPS (Schoning et al., 2005; Keita and Carfagna, 2009).

The limitations to the compass and rope method lie primarily in the burdensome nature and time required to complete the measurement. Due to the extensive time requirements, this method is often unsuitable for national level, large-sample surveys. The time requirements will of course vary by plot size but the compass and rope method is expected to be consistently and significantly more time-intensive than other objective measurement options such as GPS. However, this method is the gold standard and, as such, it is necessary to understand the sacrifices to data quality when using less time-intensive methods.

D. Remote Sensing Imagery

Satellite imagery has the potential to be used for area measurement in connection with household surveys. It is a relatively new tool which, to date, has not been largely utilized for plot area estimation but rather for land cover classification and area and yield estimation (for example, GEOSS (2009), Wardlow and Egbert (2008)). Carfagna and Gallego (2005) offer a comprehensive summary of the existing uses of satellite imagery for agricultural statistics. While this methodology has been used in conjunction with ground data, the use of remote sensing for plot area measurement has not yet been integrated into household-level surveys (to our knowledge). The use of remote sensing imagery has the potential to eliminate the need for plot visitation by identification of plot boundaries on images and subsequent measurement by geographic information systems (GIS) software, or to complement other measures for plots that are either too large or too far away for respondents and enumerators to walk to or around them. Without the need to visit each plot and conduct the measurement fieldwork time may be significantly reduced, although the time spent at the household might increase as respondents will need to identify the plot boundaries on images or maps. The effect on budget is more ambiguous. On one hand, the shortened duration of fieldwork and lighter transportation needs will reduce costs. On the other hand, additional costs may be incurred for the satellite imagery (if not relying on open sources) and post-processing of the GIS data. Despite the potential benefits of using remote sensing imagery for plot area measurement in household surveys, limitations in image resolution and availability, as well as practical implementation challenges, are anticipated to inhibit the widespread use of this method in the developing country context.

Several sources for such imagery exist with differing levels of resolution and costs. Landsat, a project executed by the US Geological Survey and NASA, has been collecting remote-sensing land data for over 40 years and is available free of charge. Its latest satellite launch, Landsat 8, captures images with pixel sizes of 15 – 30m.[7] In countries where a significant portion of the plots are less than one acre, resolution of only 30 meters is likely to result in significant measurement bias and difficultly in identifying the plot boundaries. SPOT, operated by Airbus Defence and Space, offers images of up to 2.5 meter resolution but at a significant cost (up to €8,700 for one 60km x 60km scene).[8] DigitalGlobe's Quickbird satellite collects panchromatic images with up to 61cm resolution (multispectral 2.44m resolution).[9]

The challenges that come with using aerial imagery in household surveys differ from those encountered when estimating large crop area or land cover. In order for imagery to be a successful tool in measuring plot area, one first needs to be able to distinguish the plot boundaries. This requires both cloud-free high-resolution imagery and the ability of the farmer to identify his or her plot. Additional complicating factors include tree canopy obscuring the plot boundaries and the parcel-plot concept in which parcel boundaries may be observable but plot boundaries are less clearly defined. The timing of image acquisition may also be an issue for plots that are not permanent or continuously cultivated in the same size and/or shape.

Assuming the needed imagery is available, at the desired resolution, there are several (admittedly untested) foreseen practical challenges in implementation. If respondents have limited exposure to photography and other technologies, presenting an image taken from a satellite and having them simply point to all of their agricultural plots is not likely to be an easy task. Even without consideration for plot boundary definition, the respondents may have difficulty finding the location of the plot, especially for plots that are located away from the household.

If the location of the plot is determined, outlining the boundaries may be even more difficult. If under dense canopy, boundaries will not be visible. The ability to measure plots in woody terrain or in an area common for tree crops, therefore, is impaired. In some areas, plots are nested immediately next to neighboring plots, either managed by the same household or not. If the barrier between the two plots is no more than a row of fencing crops or a narrow footpath, distinction between the two will be extremely difficult in the images. Finally, in many agricultural contexts there exists a parcel, a larger piece of land often of more permanent dimensions, and within that parcel several contiguous plots. The cropping patterns and dimensions of the plots can vary by season. If the images that the farmer is presented with do not match the current plot dimensions, identifying the boundaries will not be possible (or reliable).

Little research is available on the implementation of using remote sensing imagery for area measurement in household surveys. As technology advances and image resolution improves along with affordability,

[7] http://landsat.usgs.gov
[8] http://www2.astrium-geo.com/files/pmedia/public/r146_9_pricelist_spot_en_2012.pdf
[9] http://www.digitalglobe.com/sites/default/files/QuickBird-DS-QB-PROD.pdf

the use of this method becomes more feasible, and is likely to hold promise particularly for the measurement of large plots. Future research on the implementation of this method, including fieldwork challenges and respondent ability to identify plots, is highly encouraged. The empirical section of this paper will therefore not include a comparison of remote sensing to the other methods discussed earlier, although we plan to tackle that in future research.

III. Data: The LSMS Methodological Validation Program (MVP)

The existing literature leaves many questions to be answered about use of the various measurement methods in the household survey context. It is evident that self-reported estimates of area have an advantage in terms of time and cost, but the quality of data collected is biased systematically. The use of GPS technology is the frontrunner for objective measurement based on time requirements but the compass and rope technique is historically known for the highest accuracy.

How does GPS stack up to compass and rope under conditions when one measurement is thought to be impaired (for example, under dense canopy, in unfavorable weather, or at steep slopes)? How small is too small for GPS to be an appropriate substitute for compass and rope measurement? How can enumerator qualities affect the accuracy of measurement? Empirical evidence suggests that self-reported estimates of area are systematically biased, but under which conditions are they most accurate and how can they be improved upon? And ultimately, how does the measurement methodology employed affect the potential policy-informing conclusions drawn from analysis of the data?

In order to address the gaps in the area measurement literature and extend the applicability of studies to the plot conditions common to developing countries, the Living Standard Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) of the World Bank has prioritized land area measurement in its research agenda. The Global Strategy to Improve Agricultural and Rural Statistics has also identified improving the measurement of crop productivity, and by necessity farm area measurement, as a top priority. Because agricultural statistics are often marred by controversy over methods and overall quality, stringent validation of the available measurement methodologies is essential. With financial support from UK Aid, the Living Standards Measurement Study (LSMS) has partnered with national statistical offices in the design and implementation of methodological validation studies. The methodological studies completed through the LSMS have a particular focus on the feasibility of implementation in large-scale household surveys, thus the recommendations on best practices are made in consideration for both the highest data quality and practicality of implementation under the constraints common in this type of surveys.

To date, three methodological validation studies in Tanzania (Zanzibar), Ethiopia and Nigeria have incorporated area measurement as a fundamental component of the project. In an effort to maximize the value from each study, the land area component was complemented by at least one other aspect of the methodological research agenda. This paper uses the data collected within these studies, which are briefly described in what follows.

Data for Zanzibar, Tanzania come from the Measuring Cassava Productivity (MCP) study. The MCP focused on testing several methods for measuring cassava production, including crop-cutting, harvest diary, various means of assisted harvest diary, and various recall periods. To complement the measurement of cassava production, cassava plots were also measured using three methods of area measurement. Fieldwork extended from June 2013 – May 2014, with area measurement completed from August 2013 – January 2014. The study was conducted in two districts, one on Unguja and one on Pemba Island, Zanzibar. The sample consisted of 1247 households, with 1932 cassava plots measured for area. Partners in the study included the Ministry of Agriculture and Natural Resources, Zanzibar, the Office of the Chief Government Statistician, Zanzibar, and the World Bank. The handheld GPS Unit used in the study was a Garmin eTrex 30.

The second dataset used in this paper comes from the Ethiopia Land and Soil Experimental Research (LASER) study. The LASER study involved methodological validation of plot area measurement, soil fertility testing, and measurement of maize production. Area measurement and soil fertility testing was conducted on up to two randomly selected plots per household (where applicable, one pure-stand maize plot was selected for crop-cutting). The questionnaires were administered using computer-assisted personal interviewing. Fieldwork was conducted in multiple waves. Post-planting activities were conducted from September – December 2013. Post-harvest activities were conducted from January – early March 2014. Crop-cutting was conducted at any point during this period when the maize was deemed ready for harvest by the respondent. Area measurement was conducted in the post-planting visit. The data collection for LASER was conducted in 3 zones of the Oromia region in Ethiopia. In total, 85 enumeration areas (EAs) were randomly selected using the Central Statistical Agency of Ethiopia's Agricultural Sample Survey (AgSS) as the sampling frame. Within each EA, 12 households were randomly selected from the AgSS household listing completed September 2013. Partners in the study include the Central Statistical Agency of Ethiopia, the World Agroforestry Centre (ICRAF), and the World Bank. This study also used Garmin eTrex 30 units to collect GPS land area data.

The last batch of data used in this analysis comes from the Nigeria Area Measurement Validation Study. The primary focus of this study was validation of area measurements. Extraordinary differences between GPS and farmer self-reported areas were observed in the first wave of the Nigeria General Household Survey, igniting the need to validate the methodologies. This study was conducted on a subsample of the General Household Survey panel households. After the measurement conducted in the second wave of the national survey a special team was deployed to re-measure a subsample of plots using three area measurement methods. Fieldwork for the area measurement validation study ran from March – May 2013. Four states were selected for inclusion in the study based on safety, location, and previous performance of farmer self-reported area and GPS area. The plot selection was stratified on plot size to ensure a complete range of plot sizes included. In total, 211 households were selected, including 518 plots. The study was implemented by the National Bureau of Statistics, Nigeria, and the World Bank. The GPS Unit utilized for this study was the Garmin GPS Maps 62.

In each of the three studies, agricultural plots were measured first by farmer self-reported estimate, then by compass and rope, and finally by GPS. The order of measurements was deliberate and great attention was paid to this in the field. Farmer estimation must be recorded prior to any objective measurement so as not to influence the farmer. Enumerators were instructed in all studies *not* to influence the farmer's estimate.

Although in each of the three studies the training of methods was conducted in the same way, with the same LSMS staff present at each training, there were two differences in implementation worth noting. First, in each survey enumerators were required to repeat the compass and rope measurement if the closing error was 5% or more. However, in the Ethiopia experiment the closing error calculation was done on the spot by the enumerators, possible because of the use of computer-assisted personal interviewing. In Tanzania and Nigeria, the closing error was calculated by the supervisors (in some instances the supervisors in Nigeria were present at the plot at the time of measurement). This may influence the comparison of compass and rope with GPS measurement if, for example, an enumerator had to re-visit the plot in order to take the new compass and rope measurement and unintentionally identified the borders differently during the second visit. In Ethiopia, on the other hand, all re-measurements were completed at the same time of the initial measurement as the closing error was calculated before leaving the plot. The instance of closing error greater than 5%, however, was rare and therefore not expected to influence the analysis (in the Ethiopia experiment, only 5% of fields were measured more than once).

Second, the skill level of the enumerators varied. In Tanzania, the enumerators were the agricultural extension officers for the local area. These enumerators were very familiar with the agricultural practices but generally inexperienced in survey administration and the use of the particular measurement tools. These enumerators also happened to be significantly older and several had poor vision (requiring the purchase of glasses in order to read the compass and GPS). Because of the ongoing and intensive nature of the cassava measurement component of the Tanzania experiment, the extension officers were the preferred enumerators for the existing infrastructure and established relationships within the community. In Ethiopia, professional enumerators were hired based on past performance with the Central Statistical Agency and previous experience with computer-assisted personal interviewing (meaning some degree of familiarity with technology). In this particular study the enumerators all held bachelor's degrees and were relatively young. In Nigeria, staff from the head office of the National Statistics Bureau were trained and sent to the field rather than the enumerators use to conduct the national panel survey. In each of these studies the health, education, and skill level varied, as did the incentive structure and duration of fieldwork. Rather than discredit the comparability of the data collected from each of these studies, the consistency observed in the comparison of methods should lend confidence in the applicability across survey environments.

IV. Methods

The quest for high quality data collection in household surveys often requires tradeoffs. Limited resources and large samples can prohibit implementation of the gold standard, in this case, compass and rope measurement. It is important to know, however, how alternative methods stack up to this gold standard, the factors that cause deviation from the gold standard, and eventually means by which the measurement can be improved. The first step in analyzing the different methods is comparison of measurements. Existing literature suggests that farmer self-reported estimates are systematically biased, with farmers over-estimating the area of small plots and under-estimating the area of larger plots (DeGroote and Traore, 2005; Carletto et al., 2013a and 2013b). Some previous studies, including Palmegiani (2009), Keita and Carfagna (2009), Schoning et al. (2005) suggest that GPS measurement slightly underestimates the area with respect to compass and rope measurement. The following sections explore these conclusions on measurement precision, as well as the factors contributing to such precision, using the data collected through the three LSMS methodological validation studies described above.

To that end we construct four measures of deviation between the GPS and CR measures, defined as follows:

1. Bias: This is the simple difference between the GPS measure and the CR measure, in acres (GPS – CR);

2. Absolute value of bias: This is the absolute value of the difference between the GPS measure and the CR measure, in acres;
3. Relative bias: This is the simple difference between the GPS measure and the CR measure, in acres, divided by the CR measure, and multiplied by 100 so that it is expressed in percentage terms;
4. Absolute value of relative bias: This is the absolute value of the percent deviation between GPS and CR measure, as described in point 3. [10]

Although the main focus on what follows will be on the deviation of the GPS from the CR measure, we will occasionally employ measures of deviations of the self-reported (SR) from the CR measure, employing a terminology analogous to the one just described for the deviation of GPS from CR measures.[11]

The analysis will be based initially on a bivariate comparison of the means of the above variables for particular portions of the sample cross-tabulated with a broad range of variables of interest. The second part of the analysis will explore the determinants of the different measures of bias. We will estimate two main regression models, using the same specification for the four measures of bias described above.

The first model is an OLS regression specified as:

$$(1) \qquad Y_i = L_i + C_i + S_i + SAT_i + T_i + W_i + e_i$$

Where *Y* is one of the four measures of bias defined above, *L* is the measure of the plot taken using CR, *C* is the closing error of the CR measure, *S* is a vector of proxies for the shape of the plot (including the number of corners and the ratio of the perimeter/area, *SAT* is the number of satellites the GPS device was connected to at the time of measurement, *T* is a vector of dummy variables related to tree canopy cover (the reference being no canopy cover), *W* is a vector of dummy variables related to weather conditions at the time of the measurement (the reference being clear or partly cloudy sky), and *e* is a random error with the usual desirable characteristics.

To focus specifically on plots for which large deviations are observed between GPS and CR we then estimate a probit model to capture the factors likely to increase the probability that a plot be measured with a relative bias larger than ten percent (in absolute value). We estimate three versions of this model

[10] Mean difference presented as [$\frac{mean\ GPS - mean\ CR}{mean\ CR} \times 100$] at each level

[11] To limit the influence of outliers, observations which fell in the top 1% in terms of absolute value of relative bias (for either GPS vs CR or SR vs GPS) were dropped. A total of 77 observations were dropped under this guideline.

for each experiment, so as to investigate whether under- and over-estimation by large margins are driven by different factors. The model is specified as follows:

$$(2) \quad \Pr(Y_i = 1|X_i) = \Phi(X_i\beta)$$

where $X_i = (C_i, s_i, SAT, T_i, W_i)$ and Φ is the standard cumulative distribution function. In equation (2), Y_i is one of three outcomes: a plot having absolute relative bias greater than 10%; a plot having relative bias greater than 10%; a plot having relative bias smaller than -10%.

V. Results

A. Compass and rope: How golden the gold standard?

Compass and rope constitutes the 'gold standard' in area measurement, but, as any measurement, it is not immune from error. To some degree, the error associated with the compass and rope measurement is observable through the closing error. The closing error will not flag circumstances in which the corner points were properly identified, that can only be done through fieldwork supervision, but that the corners selected and the bearings taken create a full, closed polygon. The average closing error in our data across all measurement is around 2 percent, with the range going from 1.6 percent in Nigeria to 2.2 percent in Ethiopia (Table 1).

Regression analysis aimed at determining the factors that contribute to closing error is conducted is presented in Table 2. At the individual country level, plot size appears to influence closing error but with varying results. In Ethiopia, the data with the highest concentration of plots less than 0.05 acres, the plot area cubed has a significant and negative impact on closing error, while in Tanzania the linear area term is negative and significant, implying that closing error is smaller on larger plots. In Nigeria, where we have the largest plots, the linear term is positive while the squared term is negative. When aggregating the three experiments, plot size has no significant effect on the closing error. The number of corners on the plot (as measured by the number of vertices captured in the compass and rope measurement) exhibits a negative and significant coefficient in Ethiopia and the pooled data – contrary to the theory that more corners allows more room for measurement error. Tree cover proves to have little effect on the closing error, as none of the individual experiments exhibit significant coefficients. Had plots been randomly assigned to enumerators, enumerator effects could have been used to control for enumerator skill level and other idiosyncrasies, but plots were assigned primarily based on geographic proximity and thus enumerators often measured plots with very similar geographic properties, rendering enumerator effects inappropriate. Ultimately, there seems to be little evidence that closing error is systematic. This is comforting for the analysis that follows, as we move to explore systematic sources of error in other area measures, namely deviation from the CR method.

From the perspective of survey practitioners and national statistical offices, considerations about precision need to be accompanied by considerations related to timing (and hence cost). In survey implementation, time is often the scarcest and most valuable resource, and enumerators' remuneration an important budget item. The reason why the choice of method should matter for survey practitioners is compellingly conveyed by Figure 2, which shows the measurement time for GPS and compass and rope measurements by plot size classes, moving from small plots on the left to large plots on the right. Compass and rope requires significantly more time than GPS with time increasing exponentially with plot size, while the additional time required for GPS measurement for plots of the size included in these studies is negligible. In both the Ethiopia and Tanzania experiments[12], the compass and rope measurement took approximately four times that required for GPS. GPS required 13.9 minutes on average in Ethiopia, while the compass and rope measurement on the same plots required an average of 57 minutes. In Tanzania, the duration averages were 7.4 minutes and 29.3 minutes for GPS and compass and rope respectively. These findings are consistent with previous studies such as Schoning et al. (2005) and Keita and Carfagna (2009) who find that compass and rope takes approximately 3.5 times as long as GPS on average.

To put the time considerations into context, given the sample size and average measurement durations in Ethiopia, the field teams spent a total of 416 hours measuring plots with GPS (1797 plots * 13.89 minutes) and 1,707 hours measuring with compass and rope. Using GPS instead of compass and rope, therefore, saves 1,291 hours of labor – over 160 person/days (at 8 hours per day)! This estimate of time savings is for a relatively small-scale methodological experiment, savings in nationally representative household surveys would be proportionally larger. The true value comes in the ability of enumerators to complete the interviews of more households per day. With GPS measurement, enumerators are more likely to complete more than one household interview in a single day, allowing mobile field teams to spend less time in a particular area. Minimizing the amount of time required to collect quality land area data can significantly reduce costs and improve the flow of fieldwork.

B. Comparison of competing measurements

[12] Data on measurement duration is not available for Nigeria.

1. Compass and Rope vs. GPS

Before delving into the differences between subjective and objective measurements, we explore the relationship between the two primary means of objective measurement: GPS and compass and rope. In the literature, the main reservation regarding the use of GPS measurement in surveys is its performance on small plots. Furthermore, Keita and Carfagna (2009), Schoning et al. (2005), and Palmegiani (2009) all found that GPS tends on average to err on the negative side, i.e. to understate the area with respect to compass and rope.

Table 3 presents descriptive statistics on the GPS and compass and rope area measurements completed as part of the methodological studies. The data is presented in six classes of plot size as measured by compass and rope: Level one contains all the smallest plots (less than 0.05 acres) whereas level six includes the largest plots (greater than or equal to 1.25 acres).[13] Mean plot size is small in all countries, ranging from 0.38 acres in Ethiopia to 1.30 acres in Nigeria.

[13] Results for categories in which there are fewer than 20 observations are not reported. The same is true for all tables presented in the paper.

The mean difference between compass and rope and GPS measurement is very small (Table 3), whether expressed in acres or as a percentage of the compass and rope acreage. The sample mean bias in all three countries is plus or minus 0.01 acre, which translates in a 1-3 percent difference when expressed in relative terms (note that the values are not expressed in absolute value and as such negative and positive figures are averaged). Unlike previous studies, the data not exhibit any clear trends in terms of GPS underestimating plot size compared to CR, neither on average, nor across the distribution of plot sizes. In Nigeria, GPS under-estimates plot size compared to CR, but only slightly, while in Tanzania and Ethiopia GPS averages are somewhat larger than compass and rope measurements. Moreover, the magnitude as well as the sign of the error seem both to be unrelated to plot size, being small in all of the plot size classes.

The concern of GPS accuracy at small plot levels is also discounted. While some literature suggests that plots smaller than 0.5 hectares (1.24 acres) have significantly different GPS and compass and rope measurements with much lower correlation (Schoning et al., 2005), results from the methodological validation experiments suggest otherwise. In the pooled data, the difference between the average GPS measurement and average compass and rope measurement for plots ranging from 0.05 – 0.15 acres was less than 0.001 acres or 3% of the average compass and rope area. Even for the smallest plots, those less than 0.05 acres (202.3 square meters or 0.02 hectares), the average measurements are extremely consistent. In Ethiopia, the average GPS measurement of 390 plots in this size range is 0.0216 while the average compass and rope measurement for the same plots is 0.0215 acres.

The differences that are recorded do not appear to bear any clear trend with plot size. In Tanzania, the smallest and largest plot classes have the smallest and largest average relative bias, but figures are not large, and the number of observations in these two classes fairly small. The correlation coefficients between GPS and CR are in excess of 0.99 in all three studies, and 0.87 or larger in all classes with n larger than 50 (Table 4).

The result presented here suggest that *average* GPS measures are not much different from compass and rope even for very small pots, and even from fairly small n, and that is despite the difference in enumerator skill levels and plot characteristics of the different studies. This is confirmed by an inspection of the scatter plots in the left side of Figure 3, where GPS measures are plotted against compass and rope with measures tightly clustered around the equality line. This all lends support to the argument that GPS is an acceptable substitute of compass and rope measures across the range of plot sizes in our samples, at least if the goal is that of estimating average plot size for groups with sufficient numerosity.

2. Compass and Rope vs. Self-Reported Estimations

With an understanding of the comparability of GPS and compass and rope objective measurements, we now explore the difference in subjective (self-reported) and objective (CR) measurement[14]. As discussed above, farmer self-reported estimates of area have long been used in household surveys for the convenience and affordability of implementation. Additionally, because there is no field visit required it is easier to collect self-reported data on all plots. GPS, on the other hand is often associated with a non-negligible percentage of missing values when survey operations are large. We are not aware of large scale household surveys that use compass and rope. The benefits of subjective measurement come at a significant cost to data quality, however, as they are subject to several potential sources of error.

Table **Error! Reference source not found.**5 presents mean plot areas as measured by farmer self-reported estimation and compass and rope for all three methodological experiments. The data is grouped by compass and rope plot size class. The comparability of subjective and objective measurements is immediately questioned. While the mean plot areas as measured by GPS and compass and rope differ by only as much as 3% on average, the mean self-reported and compass and rope measurements differ by as much as 143% on average (Tanzania). The mean difference is smaller in Ethiopia and Nigeria, at 23% and 5% respectively, but still considerably larger than the divergence observed between the objective measurements.

Self-reported measures result not only in higher average deviations, but in dramatically systematic error as the size of small plots is overestimated by anywhere from 30% (Nigeria) up to a factor of six (Tanzania), with the over-estimation declining almost monotonically as plot size increases and eventually results in under-estimation in the larger plot size classes in Nigeria and Ethiopia. These results comparing objective and subjective area measures are in line with what we know from previous literature (Carletto et al. 2013a; de Groote and Traore, 2005). The scatter plots on the right side of Figure 3 convey the same message in graphic form.

Unlike the comparison the GPS and compass and rope measurements, the descriptive statistics comparing compass and rope and self-reported estimates in Table 5 exhibit consistent, clear trends in the direction of over- and under-estimation. In each of the methodological studies the area of the smaller plots is severely over-estimated. In Ethiopia for example, for plots less than 0.05 acres (as measured by compass and rope) the average farmer estimate of area is 0.09 acres, compared to an average compass and rope measurement

[14] Comparing self-reported to GPS yields exactly the same results, even when analyzed using data from the nationally representative LSMS surveys in Malawi (2010/11) and Tanzania (2010/11), therefore we limit the comparison to self-reported and CR.

of 0.02 acres (307% over-estimation on average). In Tanzania the over-estimation of small plots is even more pronounced, with an average difference in means for the smallest plots 0.32 acres, or 661% over-estimation with respect to compass and rope. While the smallest plots are significantly over-estimated, the degree of over-estimation falls with plot size and eventually, in Ethiopia and Nigeria, the sign of the bias is reversed with farmers on average *under*-estimating the area of the larger plots. This is a matter of concern not only because of the magnitude of the observed bias, but also because the bias is correlated with key variables of interest (such as plot size itself) which will translate into biases at the analysis stage (e.g. in analyses of productivity or land distribution).

Farmer over-estimation of small plots and under-estimation of larger plots is not a phenomenon observed only in the controlled environments of the methodological studies. Nationally representative data from the LSMS-ISA Malawi and Tanzania, both 2010/11, exhibit the same trends in comparison of GPS and self-reported area measurement (compass and rope not available), with the self-reported estimates on the smallest plots over-reported by more than 300% (available from the authors). Carletto et al. (2013a and 2013b) observe the same trend using LSMS-ISA data from Uganda and Niger, in addition to Malawi and Tanzania. Taken together these results point to the validity of these conclusions beyond the experiments, and certainly in much of the African continent. More studies will be needed to extend the results to other developing regions.

3. GPS measures: Exploring the deviations from the gold-standard

Previous studies have also raised the issue of how factors other than plot size may affect the quality of GPS measures, as it was recalled earlier in this paper. None of the studies have provided compelling, conclusive evidence on the impact of these factors on measurement quality. Some have explicitly called for further research to systematically investigate this matter. Our data allow us to do that for a number of factors including plot shape, slope, and tree cover, weather conditions, and number of GPS satellites acquired at the time of measurement, via a comparison of the GPS measurement to the 'gold standard' of CR measurement. While we do not go into much detail analyzing measurement error for self-reported area estimation, we do investigate the role of measurement units farmers report in (traditional vs. standard) as that is one decision that is often taken in survey design according to taste rather than based on any specific assessment of what works best.

The global navigation system requires, at a minimum, the acquisition of four satellites to triangulate the 3D position of the GPS receiver. The acquisition of additional satellites can improve position error.[15]

[15] http://www8.garmin.com/aboutGPS/

Enumerators in both the Tanzania and Ethiopia experiments recorded the number of satellites fixed at the start of the GPS measurement. In training, enumerators were instructed to wait until at least four satellites were acquired, with further instruction that they must wait until the "GPS accuracy" figure on the GPS device stabilized (therefore allowing time for maximum satellite acquisition). The mean GPS and compass and rope measurements at three levels of satellite acquisition (less than or equal to 15 satellites, 15 to 19 satellites, and more than 19 satellites) are reported in Table 6. As expected, the differences in area measurements tend to decline the higher the number of satellites, even though average differences remain small across all groups. In Ethiopia, the difference between measurements is 1.6% on the plots with fewest satellites and 1.2% on plots with the most satellites, although the trend is not linear as the middle category has an average of 1.8% bias. In Tanzania, the differences are 2.9% and 2.6%, respectively.

Various geographic and atmospheric conditions can impact the satellite acquisition. Dense canopy cover and weather conditions at the time of measurement have been found or argued to impair the precision of the GPS measurement. To address the concern over canopy density and the impact on GPS area measurement accuracy, the methodological validation studies included a subjective measure of canopy density. Table 7 presents the average GPS and compass and rope measurements disaggregated by level of tree cover. Contrary to expectations, the relative difference between the two measurements was slightly higher on plots with *no* tree cover, with the level of bias decreasing with increasing canopy density. This could be attributable to plot size, enumerator characteristics or other factors, which are not controlled for in these simple descriptive statistics. The sections below will further explore the influence of tree cover on measurement.

Higher level atmospheric conditions which can impact the satellite signals influence the precision of GPS point estimates. For this reason, the literature on GPS measurements points to weather conditions as a potential source of error. In order to address this discrepancy, all three methodological studies included a subjective measure of weather at the time of measurement, ranging from all clear to rainy. Table 8 presents the mean area measurements and the associated discrepancy by weather condition. In this case, no clear trend emerges in terms of systematic association of the differences between the two measures and weather conditions. In Ethiopia, the relative bias in measurements hovers around 2% for plots measured in conditions "mostly clear" or better. The descriptive statistics from Nigeria and Tanzania provide little evidence that weather conditions have an adverse effect on GPS area measurement. It should be noted that the majority of plots were measured in conditions "partly cloudy" or clearer.

Keita and Carfagna (2010) and Muwanga-Zake (1985) explain that plot slope can influence the difference between GPS and compass and rope measured areas, as the GPS measures the horizontal plane and traversing measures the surface area. Freemont and Benson (2011) note that plot slopes greater than 10 degrees will result in significantly different measurements. The LASER study incorporated the use of clinometers for slope measurement. Descriptive statistics on the slope and measurement bias are reported in Table 9. For plots of slope less 5 degrees or less, the mean relative difference is 1.3% whereas for plots of 6 – 15 degrees it is 2.2% and for plots of slope greater than 15 degrees it is 2.4%. Plot slope may play a role in the discrepancy observed between GPS and compass and rope area measurement, but the materiality of the impact is questionable.

Having ascertained that the average difference between GPS and CR is small does not rule out that for individual measurements, there may be observations with errors of significant magnitude. To investigate this aspect we plot the percentage and absolute differences between GPS and CR measures over plot area (Figure 4). A number of considerations emerge from a visual analysis of these graphs. First, the GPS measurement error in percentage terms is often far from negligible, in some instance larger than plus/minus 50 percent. Second, large percentage errors appear to be roughly equally distributed above or below the zero line, which explains why we do not observe differences in the means for the two measures. Thirdly, the magnitude of the percentage errors is much larger for the small size classes, and decreases rapidly as plot size increase. Those trends are clearly mirrored by the graphs with the absolute bias, which show no clear correlation with plot size and fairly constant dispersion both sides of the zero line, with most values within the plus/minus 5 range. That seems to suggest that it is the inherent imprecision of GPS devices that causes percentage error to matter much more for very small plots. We therefore turn to investigating more in depth the extent and nature of the errors for observation with an arbitrary set value of plus/minus 10 percent.

Table 10 reports on some key the characteristics of the plots and the measurements, slicing the sample according to whether the bias is below or above the 10 percent threshold, and splitting the latter portion of the sample in observations where GPS over- or under-reports land area. The first thing to observe is that the number of observations with such large errors is far from negligible, ranging from 17 percent in Nigeria to 31 percent in Ethiopia. Again, no strong systematic bias emerges in terms of GPS over- or under-reporting: differences in average acreage between GPS and CR are small even for the high-error portion of the sample in all countries. In two of the three experiments (Ethiopia and Tanzania) there are more GPS observations with large percentage over-reporting compared to under-reporting, in Nigeria the opposite is true.

The table also reports the average values for several of the variables that are expected to influence the quality of GPS measurement. We do not observe any substantial difference for some of the factors that are often cited as important for GPS measurement, such as number of plot corners, number of satellites and tree canopy cover. We do however observe some difference in the perimeter/area ratio, which approximates the complexity of a plot shape. In all three experiments this is higher in plots that are substantially underestimated by the GPS measure, compared to the plots that are measured with greater accuracy. Ethiopia is the only country for which such a difference, albeit of much smaller magnitude, is also observed for plots with size over-reported by more than 10 percent. Shape complexity therefore does seem to affect GPS precision, resulting mostly in under-reporting of the plot size.

Some differences are observed also for the walking speed of the enumerators, but with results that are more difficult to interpret. In Ethiopia the walking speed is lower on the plots measured with larger errors. In Tanzania we observe no sizeable difference. In Nigeria plots that are under-estimated by more than 10 percent are associated with lower average walking speed, while plots that are over-estimated tend to record higher average walking speed. We are not able to draw any conclusions from this mixed evidence.

One last variable for which we do observe systematic differences is the magnitude of the CR closing error. In both Ethiopia and Tanzania there is a gradient between plots that are underestimated by a large margin (which have the smallest closing error), plots with error below 10 percent (which have in-between closing error), and plots with large GPS over-estimate (with the largest closing error). In Nigeria the plots with larger over-estimates are also the ones with the largest closing error, but the ranking of the other two groups is inverted. What we conclude from these observations is that for the cases in which we observe substantial deviations between GPS and CR measures, part of the explanation is likely to rest in noise in the CR measures. In that sense the inaccuracy in the GPS measures may be somewhat less serious than if one just looked at the prevalence of high bias cases, and that as observed earlier the gold standard is also bound to be imperfect.

C. Regression analysis of the differences between competing measures

1. Pulling them all together. A multivariate analysis of factors contributing to bias between CR and GPS

The descriptive statistics presented above are aimed at comparing the two primary objective area measurement options, GPS and compass and rope. In this section, regression analysis is used to explore the determinants of measurement bias, defined here as the difference between GPS measured area and compass and rope measured area, which is used as the benchmark.

The results in Table 11 include four specifications per dataset, the difference among them being the dependent variable, which is: (i) bias (GPS – CR), (ii) absolute value of bias, (iii) relative bias (bias divided by CR area * 100%), and (iv) absolute value of relative bias. Recall from the descriptive statistics that the observed error is generally small, and that we found little evidence of systematic variation with many of the factors that are *a* priori expected to influence GPS measurement precision. It is therefore not completely surprising that the explanatory power of these regressions (as captured by their R^2 values) is low, and that a majority of the estimated coefficients are not statistically significant.

The main variables of interests are the set of terms (levels, quadratic, cubic) related to the plot size itself, as measured by CR, graphic representations of which are available in Figure 5. In the first specification, there appears to be a relationship between plot size and measurement error only in Ethiopia, where the shape of the relationship is that of an inverted U, with the predicted bias being positive on very small plots, peaking at about 0.7 acres, and becoming negative for plots larger than about 1.7 acres. The coefficients are small, so that the predicted error is in the plus/minus 0.02 acres range. In Tanzania, a linear relationship is exhibited in which larger plot size results in larger bias (in terms of acres). In Nigeria and the pooled data there is no statistically significant relationship between bias and plot size, controlling for other factors.

When the absolute level is considered, in the second specification, the relationship with plot size becomes monotonically positive, with a small curvature (quadratic term is significant) only in Nigeria. Values are somewhat larger, up to about 0.2 acres in the observed plot size range, but still small.

When the percentage bias is considered (third specification) the relationship with plot size becomes an L-shaped quadratic curve (the cubic term is significant only in Ethiopia and the pooled data, tilting the curve up around the 2 acres mark). The last specification has the absolute bias expressed in percentage terms as the dependent variable, with the relationship with plot size being again best characterized as L-shaped.

Taken together these results are in line with the earlier descriptive analysis in that the overall distribution of the bias does not seem to bear much relationship with plot size, as they are largely equally distributed on the positive and negative side. The absolute magnitude of the error does however increase somewhat with plot size, but less than proportionally. For that reason, in percentage terms the bias actually declines fairly rapidly as plot size increase, stabilizing as plot size reaches the 1-2 acres range.

Of the other covariates reflecting physical characteristics expected to affect the quality of GPS measures (cloud and canopy cover, plot slope) hardly any are consistently significant across country. What appears to matter most are actually closing error, and the perimeter/area ratio. The former reflects inaccuracy in the CR measure, while the latter is a proxy for the complexity of the plot shape which is likely to affect the accuracy of GPS measures, but can in principle also be capturing noise in the CR measure besides what is captured by the closing error. In Ethiopia and Tanzania, there is evidence that heavy canopy cover does increase relative bias (in absolute value, specification 4).

Although the difference between the two objective measurements is relatively small on average, it is it worth digging into the problem cases in which the deviation is much larger. Table 12 report results of a probit model (see equation 2) estimating the probability of a plot being a "problem plot" – defined here as having a relative bias greater than 10% (in absolute value). In all countries we find evidence of a cubic relationship between the probability of GPS overestimating are by more than 10 percent and plot size. That translates into the probability being highest for very small plots, and decreasing fast as plot size increases, before flattening fairly quickly (and eventually tilting up somewhat) for larger plot sizes. The same relationship is found for plots under-estimated by GPS in Ethiopia, but not in the other two experiments. As in the OLS regression, in the probit model the other covariates that appear to be playing a role are closing error and perimeter/area ratio. Tree canopy cover appears to play more of a role in these regressions, implying that the effects of canopy cover are not felt equally throughout the distribution of the bias variable, but that they kick-in in particular regions of the distribution. Weather at the time of GPS measurement has a more limited effect. In Tanzania, plots that are measured during "mostly cloudy", "all cloudy", or "rainy" weather are slightly more likely to be over-stated by the GPS by 10% or more (compared to plots measured during "partly cloudy" or clearer weather). In all other countries and specifications the cloudy or rainy weather does not have a statistically significant effect on the probability of area being measured with high bias.

2. Comparison of SR and Objective Measurements

While Table 5 illustrates the degree to which farmer self-reported estimates differ from compass and rope measurements, it does not offer any explanation as to why the two systematically diverge. For this we turn to regression analysis. Table 13 presents the results of three OLS regression models: the first on the measurement bias (farmer self-reported estimate minus CR measured area), the second on the absolute value of this bias, and the third on the absolute value of the relative bias (in percentage terms). Results are presented for methodological studies in Ethiopia and Nigeria.

The claim of plot area affecting the direction and degree of error associated with self-reported area estimates is supported by the regression results. In both studies, in the first specification (on bias) the coefficient on quadratic plot area term is negative and positive in the cubic term. In the second specification (on absolute value of bias), the coefficients on plot area are positive suggesting that as plot size increases the degree of farmer over-reporting shrinks while at the same time the absolute value of the bias increases. When looking at the absolute value of relative bias, the linear term is negative and the quadratic term positive in both studies but at very different magnitudes, potentially driven by the difference in average plot size observed across the two countries.

The distance from the plot to the dwelling also holds significant explanatory power. The results from Ethiopia suggest that self-reported estimates of area diverge more from compass and rope measurements on those plots that are further from the household. This could be theoretically explained by assuming the farmer spends less time on plots more distant from the household and does not have the opportunity to view these plots to make his/her area estimate, should he/she prefer to do so. Consistent with Carletto et al. (2013a), the existence of property rights (proxied here by the possession of a title or certificate of ownership or the ability to sell or use the plot as collateral) has a significant, negative relationship with the relative bias in Ethiopia, suggesting that on plots where the household has some form of property right they are better able to estimate the area.

Household characteristics such as the gender, age, and education of the household head present differently across countries. Results from Nigeria (but not Ethiopia) suggest that measurement bias is greater in households with older household heads. Contrary to expectations, the education and literacy status of the household head does not hold consistent results across country. In Ethiopia, literacy of the household head reduces measurement bias (in acreage terms), but years of education has the opposite. In fact, the education of the household head significantly increases the relative bias in Ethiopia (but is not statistically significant in Nigeria).

VI. Conclusions

There are several important findings emerging forcefully from this analysis, which translate into clear implications for future survey design and implementation. The first result is that our experimental data confirm what we already knew about the magnitude, direction and determinants of measurement error in farmers' self-reported estimates of land area. While not a novel finding, this is a useful reminder of the urgency to find alternative measures that are both accurate and usable in the context of large scale household surveys. GPS measurement is the obvious candidate.

Much of the focus of the paper has, therefore, been on assessing the fitness for purpose of GPS measures. In this respect an important finding of the study is that on average GPS measures return very accurate estimates of plot size, even for very small plots, and even for reasonably small samples. We also do not detect any evidence that GPS systematically under-report land size, as is the case in earlier studies. That should suffice to make GPS an attractive method for land data collection for most household survey practitioners. This conclusion becomes even more forceful when taken together with the comparison of the time (cost) of GPS compared to CR measurement, with our data showing GPS to lead CR by several orders of magnitude.

This strong message in support of the adoption of GPS in survey fieldwork, is, however, mediated by a number of considerations regarding outstanding challenges with GPS measurement. One that emerges from the analysis is that while the GPS measurement error is almost universally small in magnitude (only 5 percent of observations recording a discrepancy with CR of more than 0.09 acres), in relative terms a discrepancy of plus/minus 10 percent is not uncommon. Considering that GPS measures in large scale surveys are often plagued by a missingness rate in the range of 15-30 percent, that means that a large scale dataset of GPS plot measurements can be plagued by as much as 50 percent problem cases. That requires complementing data collection with information that can aid in identifying those problem cases, as well as with thinking of field implementation protocols that can help reduce both bias and missingness in GPS measures.

Despite the evidence that subjective measurements can be riddled with problems farmer self-reported estimates of area should still be included in household surveys, but not as the primary measurement method. Objective measurements come with their own challenges, including time and equipment requirements, questions of accuracy at small-plot levels, and feasibility of full plot sample measurement. Subjective measurements have negligible fieldwork costs, and, more importantly, they can serve as a baseline for imputation where objective measurements may be missing (Kilic et al. 2013). Therefore, we recommend GPS measurement (where feasible) complemented by farmer self-reported estimated area (for all plots).

Another ancillary story emerging from the data concerns self-reported data. While confirming all the known issues with measurement error in self-reported land data, the analysis presented here provides at least one suggestion for limiting the scope of this error in the future. When land area is collected using non-standard units, as opposed to forcing respondents or enumerators to perform a conversion to standard units at the time of the interview, data from self-report appears to approximate our preferred measures (GPS and CR) much better. The implication for survey work is to allow the reporting in non-standard units in questionnaires, while focusing on collecting good conversion factors.

Finally, our analysis casts some shadows on the benchmark compass and rope measurement. It appears that a good deal of what we labeled for simplicity as GPS measurement error, may in fact be linked to noise in the CR data. This is hardly surprising since CR does in fact require a good deal of precision that, no matter how careful training, will be hard to reach for survey enumerators that are not professional land surveyors. In terms of specific suggestions for CR measurements we do observe an increase in discrepancy between CR and GPS when the CR closing error is above 3 percent. Translated in recommendations for survey work, this means that 3 percent offers a good rule of thumb for instructing enumerators to re-take CR measurement.

VII. References

Andre, C. and Platteau, J.-P. (1998). "Land Relations Under Unbearable Stress: Rwanda Caught in the Malthusian Trap." *Journal of Economic Behavior & Organization*, vol. 34, pp.1 – 47.

Barker R.M and A.B. Forbes (2001). *Discrete Model Validation*. Software Support for Metrology Best Practice Guide No. 10. National Physical Laboratory, Queens Road, Teddington, Middlesex, United Kingdom.

Baten, J., and Juif, D. (2013). "A story of large landowners and math skills: Inequality and human capital formation in long-run development, 1820–2000." *Journal of Comparative Economics*, http://dx.doi.org/10.1016/j.jce.2013.11.001

Bezu, S. and Holden, S. (2013). "Unbundling Land Administrative Reform: Demand for Second Stage Land Certification in Ethiopia,"CLTS Working Papers 3/13, Centre for Land Tenure Studies, Norwegian University of Life Sciences.

Bogaert, P., J. Delincè and S. Kay. (2005). "Assessing the error of polygonal area measurements: a general formulation with applications to agriculture." Meas. Sci. Technol. 16 (2005) 1170–1178 doi:10.1088/0957-0233/16/5/017 .

Carfagna, E. and Gallego, F.J. (2005). "Using Remote Sensing for Agricultural Statistics." *International Statistical Review* , **73,** 3, 389–404.

Carletto, G., Gourlay, S., and Winters, P. (2013). "From Guesstimates to GPStimates: Land Area Measurement and Implications for Agricultural Analysis," *World Bank Policy Research Working Paper 6550*.

Carletto, G., Savastano, S., and Zezza, A. (2013). "Fact or artifact: The impact of measurement errors on the farm size–productivity relationship," *Journal of Development Economics*, Elsevier, vol. 103(C), pages 254-261.

De Groote, H. and O. Traoré. (2005). "The cost of accuracy in crop area estimation". *Agricultural Systems*. 84 (2005) 21–38.

Deininger, K., and Lyn Squire. (1998). "New ways of looking at old issues: inequality and growth." *Journal of Development Economics*, Volume 57, Issue 2, 1998, Pages 259-287.

FAO (1982). Estimation of crop areas and yields in agricultural statistics, FAO, Rome.

Fermont, A., T. Benson (2011). "Estimating yield of food crops grown by smallholder farmers: A review in the Ugandan context." *Uganda Strategy Support Program Working Paper* No. USSP 05, IFPRI.

GEOSS (2009). Best practices for crop area estimation with Remote Sensing. Edited by Gallego J., Craig M., Michaelsen J., Bossyns B., Fritz S. Ispra, June 5-6, 2008.

Graesser, J. and Long, J. (Forthcoming). A Python Library for Remote Sensing, MapPy v 0.3.5, pre-release

Hofmann-Wellenhof, B., H. Lichtenegger, E. Wasle. (2008). "GNSS – Global Navigation Satellite Systems." Springer-Verlag: New York, NY.

Kay, S. and A. Sima (2009). "Area Measurement Validation Scheme." JRC Technical Notes. Luxembourg: Office for Official Publications of the European Communities

Keita, N., and Carfagna. E. (2009). "Use of modern geo-positioning devices in agricultural censuses and surveys: Use of GPS for crop area measurement." *Bulletin of the International Statistical Institute*, the 57th Session, (2009), Proceedings, Special Topics Contributed Paper Meetings (STCPM22), Durban.

Keita, N., Carfagna, E., and Mu'Ammar, G. (2010). "Issues and guidelines for the emerging use of GPS and PDAs in agricultural statistics in developing countries." *The Fifth International Conference on Agricultural Statistics*; Kampala, Uganda.

Kelly, V. and C. Donovan (2008). "Agricultural Statistics in Sub-Saharan Africa: Differences in Institutional Arrangements and their Impacts on Agricultural Statistics Systems A Synthesis of Four Country Case Studies." *Michigan State University International Development Working Paper* No. 95, East Lansing, Michigan.

Kilic, T., Zezza, A., Carletto, G., and Savastano, S. (2013). "Missing(ness) in Action: Selectivity Bias in GPS-Based Land Area Measurements," *World Bank Policy Research Working Paper 6490*.

Macours, K. (2011). "Increasing Inequality and Civil Conflict in Nepal." *Oxford Economic Papers*, vol. 63, n. 1, pp. 1-26.

Muwanga-Zake E.S.K. and J.B. Magezi-Apuuli. (2005). Experience with GPS Equipment in Measuring Crop areas: The Case of Uganda. *African Statistical Journal*. Vol1.

Magezi-Apuuli, J., E. Menyha, E. Muwanga-Zake, P. Schoning (2005). "GPS equipment for agricultural statistics surveys - lessons learned from fieldwork in Uganda." Retrieved on 06/13/2011 from http://www.nass.usda.gov/mexsai/Papers/gpsequipmentp.pdf.

Muwanga-Zake, E. S. K. (1985). "Sources of Possible Errors and Biases in Agricultural Statistics in Uganda: A Review." Kampala, Uganda: Institute of Statistics and Applied Economics, Makerere University.

Palmegiani, G. (2009). Measuring cultivation parcels with GPS: a statistical evidence. Mimeo, LUISS University: Rome.

Schoning, P., Apuuli, J. B. M., Menyha, E., Zake-Muwanga, E. S. K. (2005). "Handheld GPS equipment for agricultural statistics surveys: Experiments on area-measurement and geo-referencing of holdings done during fieldwork for the Uganda Pilot Census of Agriculture." Statistics Norway Report 2005/29.

Wardlow, B., and Egbert, S. (2008). "Large-Area Crop Mapping Using Time-Series MODIS 250m NDVI Data: An Assessment for the U.S. Central Great Plains." Remote Sensing of Environment, 112, 1096–1116.

________ (2009). "How to Feed the World in 2050. High Level Expert Forum – The Special Challenge for Sub-Saharan Africa." Rome, Italy: FAO.

________ (2010). *Rural poverty report 2011 New realities, New challenges : New opportunities for tomorrow's generation*. Rome, Italy: IFAD.

VIII. Tables

Table 1 - Closing Error and Plot Shape

Plot Shape	Ethiopia			Tanzania			Nigeria			Pooled		
	N	Closing Error (%)	Bias (GPS-CR, acres)	N	Closing Error (%)	Bias (GPS-CR, acres)	N	Closing Error (%)	Bias (GPS-CR)	N	Closing Error (%)	Bias (GPS-CR, acres)
< = 4 sides	662	2.32	0.01	358	2.03	0.00	71	1.46	-0.01	1091	2.17	0.01
5 - 9 sides	875	2.19	0.00	980	2.00	0.01	215	1.63	-0.01	2070	2.04	0.00
> = 10 sides	228	2.09	0.01	570	1.97	0.01	200	1.72	-0.01	998	1.95	0.01
Total	1765	2.23	0.01	1908	2.00	0.01	486	1.64	-0.01	4159	2.05	0.01

Table 2 – Determinants of Closing Error

OLS Regression

	Ethiopia	Tanzania	Nigeria	Pooled
CR Area (acres)	0.02	-0.291**	0.132*	-0.023
CR Area^2	0.166	-	-0.007*	-
CR Area^3	-0.039**	-	-	-
Number of Corners	-0.032***	0.006	-0.006	-0.014***
Slope (clinometer)	0.014***	-	-	-
Treecover:				
Partial	-0.003	-0.052	-0.104	-0.139***
Heavy	0.183	-0.163	0.324	0.020
Weather:				
Mostly Cloudy - Rainy	-0.014	-0.128**	0.021	-0.05
Constant	2.298***	2.118***	1.586***	2.237***
N	1765	1908	486	4159
R2	0.015	0.007	0.019	0.008

*p<.1; ** p<.05; *** p<.01

Table 3 – GPS vs Compass and Rope (CR) measures, by plot size classes

Means; Acres

Level (CR)	Ethiopia				Tanzania				Nigeria				Pooled			
	N	GPS	CR	Mean Bias / Mean CR	N	GPS	CR	Mean Bias / Mean CR	N	GPS	CR	Mean Bias / Mean CR	N	GPS	CR	Mean Bias / Mean CR
1 (< 0.05 acres)	390	0.02	0.02	0%	45	0.04	0.04	-3%	-	-	-	-	436	0.02	0.02	-1%
2 (< 0.15 acres)	400	0.10	0.09	2%	631	0.11	0.11	2%	21	0.11	0.11	-7%	1052	0.10	0.10	2%
3 (< 0.35 acres)	365	0.24	0.24	3%	823	0.23	0.23	2%	73	0.24	0.25	-4%	1261	0.24	0.23	2%
4 (< 0.75 acres)	328	0.52	0.51	2%	326	0.51	0.49	4%	129	0.52	0.53	-2%	783	0.51	0.50	2%
5 (< 1.25 acres)	182	0.98	0.96	2%	63	0.94	0.92	2%	108	0.97	0.99	-2%	353	0.97	0.96	1%
6 (>= 1.25 acres)	100	1.91	1.89	1%	20	1.91	1.81	5%	154	2.87	2.87	0%	274	2.45	2.44	1%
Total	1765	0.38	0.38	2%	1908	0.28	0.27	3%	486	1.30	1.31	-1%	4159	0.44	0.44	1%

Table 4 – Correlation Coefficient (GPS & CR)

Level (CR)	Ethiopia	Tanzania	Nigeria	Pooled
1 (< 0.05 acres)	0.95	0.81	-	0.95
2 (< 0.15 acres)	0.91	0.92	0.91	0.92
3 (< 0.35 acres)	0.90	0.95	0.90	0.93
4 (< 0.75 acres)	0.91	0.96	0.87	0.92
5 (< 1.25 acres)	0.93	0.95	0.91	0.92
6 (>= 1.25 acres)	0.98	0.96	1.00	0.99
Total	0.996	0.993	0.997	0.997

Acres

Level (CR)	Ethiopia N	Ethiopia SR	Ethiopia CR	Ethiopia Mean Bias / Mean CR	Tanzania N	Tanzania SR	Tanzania CR	Tanzania Mean Bias / Mean CR	Nigeria N	Nigeria SR	Nigeria CR	Nigeria Mean Bias / Mean CR	Pooled N	Pooled SR	Pooled CR	Pooled Mean Bias / Mean CR
1 (< 0.05 acres)	352	0.09	0.02	307%	44	0.32	0.04	661%	-	-	-	-	397	0.12	0.03	371%
2 (< 0.15 acres)	392	0.27	0.09	188%	622	0.41	0.11	288%	21	0.15	0.11	30%	1035	0.35	0.10	247%
3 (< 0.35 acres)	351	0.40	0.23	72%	816	0.62	0.23	173%	73	0.39	0.25	55%	1240	0.55	0.23	136%
4 (< 0.75 acres)	316	0.66	0.51	29%	323	0.98	0.49	100%	129	0.79	0.53	50%	768	0.82	0.50	62%
5 (< 1.25 acres)	179	0.95	0.97	-2%	63	1.53	0.92	66%	108	1.31	0.99	33%	350	1.16	0.96	21%
6 (>= 1.25 acres)	99	1.42	1.90	-25%	20	2.05	1.81	13%	154	2.57	2.87	-10%	273	2.11	2.44	-13%
Total	1689	0.47	0.38	23%	1888	0.65	0.27	143%	486	1.38	1.31	5%	4063	0.67	0.44	51%

Table 5 – Comparison of Self-Reported and CR measures

Table 6 – Bias and number of satellites

Acres

Number of Satellites	Ethiopia N	Ethiopia GPS	Ethiopia CR	Ethiopia Bias	Ethiopia Mean Bias / Mean CR	Tanzania N	Tanzania GPS	Tanzania CR	Tanzania Bias	Tanzania Mean Bias / Mean CR
<= 15	305	0.23	0.22	0.00	1.6%	488	0.31	0.30	0.01	2.9%
16 - 19	1204	0.40	0.39	0.01	1.8%	1266	0.27	0.26	0.01	2.9%
> = 20	256	0.49	0.48	0.01	1.2%	154	0.24	0.24	0.01	2.6%
Total	1765	0.38	0.38	0.01	1.7%	1908	0.28	0.27	0.01	2.9%

Acres

Tree Cover	Ethiopia N	Ethiopia GPS	Ethiopia CR	Ethiopia Mean Bias / Mean CR	Tanzania N	Tanzania GPS	Tanzania CR	Tanzania Mean Bias / Mean CR	Nigeria N	Nigeria GPS	Nigeria CR	Nigeria Mean Bias / Mean CR	Pooled N	Pooled GPS	Pooled CR	Pooled Mean Bias / Mean CR
None	1245	0.40	0.39	2.0%	858	0.27	0.26	3.2%	150	0.92	0.94	-2.1%	2253	0.38	0.38	1.6%
Partial	431	0.35	0.35	0.7%	946	0.28	0.27	2.8%	279	1.43	1.44	-0.5%	1656	0.49	0.49	0.8%
Heavy	89	0.28	0.28	1.0%	104	0.34	0.33	1.6%	57	1.66	1.66	-0.1%	250	0.62	0.62	0.4%
Total	1765	0.38	0.38	1.7%	1908	0.28	0.27	2.9%	486	1.30	1.31	-0.8%	4159	0.44	0.44	1.2%

Table 7 - Bias and tree cover

Acres

Weather	Ethiopia N	Ethiopia GPS	Ethiopia CR	Ethiopia Mean Bias / Mean CR	Tanzania N	Tanzania GPS	Tanzania CR	Tanzania Mean Bias / Mean CR	Nigeria N	Nigeria GPS	Nigeria CR	Nigeria Mean Bias / Mean CR	Pooled N	Pooled GPS	Pooled CR	Pooled Mean Bias / Mean CR
Clear/Sunny	576	0.50	0.49	1.9%	800	0.30	0.29	3.3%	318	1.32	1.33	-0.8%	1694	0.56	0.55	1.1%
Mostly Clear	733	0.33	0.33	2.0%	217	0.28	0.27	1.3%	101	1.35	1.34	0.5%	1051	0.42	0.41	1.5%
Partly Cloudy	334	0.32	0.31	0.6%	705	0.25	0.24	2.8%	62	1.13	1.16	-2.6%	1101	0.32	0.32	1.0%
Mostly Cloudy	94	0.34	0.34	-0.1%	76	0.22	0.21	3.1%	-	-	-	-	174	0.30	0.30	0.5%
Completely Cloudy	-	-	-	-	41	0.26	0.25	2.0%	-	-	-	-	54	0.25	0.25	1.3%
Rainy	-	-	-	-	69	0.29	0.28	2.7%	-	-	-	-	85	0.34	0.33	2.1%
Total	1765	0.38	0.38	1.7%	1908	0.28	0.27	2.9%	485	1.30	1.31	-0.7%	4159	0.44	0.44	1.2%

Table 8 – Bias and weather conditions

Table 9 – Bias and slope (measured with a clinometer)

Plot Slope (degrees)	Ethiopia				
	N	GPS	CR	Bias	Mean Bias / Mean CR
0 - 5	1076	0.36	0.35	0.00	1.3%
6 - 15	562	0.44	0.43	0.01	2.2%
> 15	127	0.37	0.36	0.01	2.4%
Total	1765	0.38	0.38	0.01	1.7%

Bias = GPS - CR (acres)

	Ethiopia					Tanzania					Nigeria					Pooled				
	Relative Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Relative Bias < 10%	All Plots	Relative Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Relative Bias < 10%	All Plots	Relative Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Relative Bias < 10%	All Plots	Relative Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Relative Bias < 10%	All Plots
N:	542	305	237	1223	1765	388	251	137	1520	1908	84	31	53	402	486	1014	587	427	3145	4159
% of Total Plot Sample	31%	17%	13%	69%	100%	20%	13%	7%	80%	100%	17%	6%	11%	83%	100%	24%	14%	10%	76%	100%
Average:																				
CR Area (acres)	0.22	0.23	0.20	0.45	0.38	0.21	0.23	0.17	0.28	0.27	0.96	1.37	0.72	1.38	1.31	0.28	0.29	0.25	0.49	0.44
GPS Area (acres)	0.23	0.28	0.16	0.45	0.38	0.22	0.27	0.14	0.29	0.28	0.99	1.64	0.61	1.37	1.30	0.29	0.35	0.21	0.49	0.44
Bias (GPS - CR)	0.01	0.05	-0.03	0.00	0.01	0.02	0.04	-0.03	0.01	0.01	0.03	0.26	-0.11	-0.02	-0.01	0.01	0.05	-0.04	0.00	0.01
|% Bias|	22.74	23.18	22.18	3.41	9.34	16.95	16.61	17.57	4.13	6.74	18.96	23.45	16.33	3.93	6.53	20.21	20.38	19.97	3.82	7.82
Closing Error (%)	2.24	2.31	2.14	2.22	2.23	2.09	2.22	1.86	1.97	2.00	2.14	2.91	1.69	1.54	1.64	2.17	2.30	1.99	2.01	2.05
Number of Corners	5.94	5.98	5.89	6.48	6.32	8.09	8.25	7.80	8.35	8.30	9.19	9.55	8.98	10.24	10.06	7.03	7.14	6.89	7.87	7.66
Per : Area Ratio (GPS)	0.41	0.27	0.59	0.20	0.26	0.19	0.16	0.24	0.15	0.16	0.12	0.08	0.15	0.09	0.09	0.30	0.21	0.42	0.16	0.20
Number of Satellites	16.9	17.1	16.8	17.4	17.3	16.5	16.6	16.4	16.7	16.7	-	-	-	-	-	-	-	-	-	-
Walking Speed (m/min)	37	37.7	36.10	43.4	41.4	42.1	42.3	41.5	44.3	43.8	58.5	66.9	53.3	61.8	61.2	40.7	41.3	39.8	45.9	44.7
Treecover:																				
Partial (n)	139	77	62	292	431	207	132	75	739	946	51	19	32	228	279	397	228	169	1259	1656
(%)	26%	25%	26%	24%	24%	53%	53%	55%	49%	50%	61%	61%	60%	57%	57%	39%	39%	40%	40%	40%
Heavy (n)	39	20	19	50	89	26	16	10	78	104	9	3	6	48	57	74	39	35	176	250
(%)	7%	7%	8%	4%	5%	7%	6%	7%	5%	5%	11%	10%	11%	12%	12%	7%	7%	8%	6%	6%

Table 10 –Descriptive statistics for high-bias observations

OLS Regression
Bias = GPS - CR (acres)

	Ethiopia				Tanzania				Nigeria				Pooled			
			{Bias/CR }	{|Bias|/CR }			{Bias/CR }	{|Bias|/CR }			{Bias/CR }	{|Bias|/CR }			{Bias/CR }	{|Bias|/CR }
Dependent Variable:	Bias	|Bias|	* 100	* 100	Bias	|Bias|	* 100	* 100	Bias	|Bias|	* 100	* 100	Bias	|Bias|	* 100	* 100
CR Area (acres)	0.049***	0.040***	-9.186***	-9.013***	0.055**	0.069***	-8.169***	-5.421***	-0.005	0.056***	-0.325	-0.011	0.005	0.052***	-3.021***	-2.141***
CR Area^2	-0.037***	-	4.025***	4.557***	-	-	3.057**	2.406**	-	-0.002***	-	-	-0.001*	-0.002***	0.462***	0.360***
CR Area^3	0.006***	-	-0.477**	-0.589***	-	-	-	-	-	-	-	-	-	-	-0.016***	-0.013***
Closing Error (%)	0.003**	0.001	0.862***	0.156	0.002*	0.002**	0.494***	0.282**	0.012	0.022***	1.405***	1.307***	0.004***	0.004***	0.916***	0.400***
Number of Corners	0.001	0.001	-0.048	0.026	0.000	0.000	0.056	0.090**	-0.001	0.000	-0.102*	-0.009	0.000	0.000	-0.014	-0.01
Per : Area Ratio (GPS)	0.002	-0.002	-13.927***	10.551***	-0.003	0.036	-38.537***	12.955***	-0.133	0.113	-58.876***	43.418***	-0.006*	0.001	-13.254***	12.254***
Number of Satellites	0.000	-0.001	0.107	-0.114	0.000	0.000*	0.027	-0.004	-	-	-	-	-	-	-	-
Slope (clinometer)	0.000	0.000	0.018	0.045	-	-	-	-	-	-	-	-	-	-	-	-
Treecover:																
Partial	-0.005**	0.000	-0.074	0.318	-0.001	0.002	-0.729*	0.779**	0.015	-0.002	0.157	0.739	-0.002	0.002	-0.711*	0.481*
Heavy	-0.006	0.009*	-1.019	3.804**	-0.006*	0.001	-1.339	1.116*	0.02	-0.014	-1.266	0.543	-0.005	0.004	-1.418	2.286***
Weather:																
Mostly Cloudy - Rainy	-0.005**	0.003	-0.085	1.206*	0.001	0.000	0.569	0.527*	-0.03	0.015	-0.29	-0.782	-0.002	0.002	0.451	0.368
Constant	-0.009	0.007	3.897	9.213***	-0.007	-0.021	8.317***	3.867**	-0.01	-0.039	2.734	0.034	-0.005	-0.009**	3.465***	4.942***
N	1765	1765	1765	1765	1908	1908	1908	1908	486	486	486	486	4159	4159	4159	4159
R2	0.046	0.262	0.117	0.199	0.151	0.263	0.071	0.051	0.027	0.323	0.134	0.14	0.014	0.319	0.086	0.152

*p<.1; ** p<.05; *** p<.01

Table 11 – Determinants of Bias (GPS – CR)

Table 12 – Determinants of High Bias

Probit (reporting marginal effects), Error Clustered on Enumerator ID
Bias = GPS - CR (acres)

	Ethiopia			**Tanzania**			**Nigeria**			**Pooled**		
Dependent Variable:	Percent Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Percent Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Percent Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%	Percent Bias > 10%	GPS over-reported by > 10%	GPS under-reported by > 10%
CR Area (acres)	-0.899***	-0.941***	-0.339***	-0.674***	-0.871***	-0.036	0.006	-0.194***	0.008	-0.215***	-0.250***	-0.074***
CR Area^2	0.794***	1.056***	0.281***	0.555**	0.642***	-	-	0.060***	-	0.059***	0.099***	0.017***
CR Area^3	-0.198***	-0.349***	-0.059**	-0.117*	-0.129**	-	-	-0.005***	-	-0.004***	-0.010***	-0.001**
Closing Error (%)	0.003	0.01	-0.007	0.014*	0.018**	-0.005	0.055***	0.039***	0.014**	0.020***	0.023***	-0.005
Number of Corners	0.006	0.002	0.004	0.005	0.006*	0.001	-0.001	-0.001	0.000	0.000	0.000	0.000
Per : Area Ratio (GPS)	0.173***	-0.144**	0.175***	0.428***	-0.911***	0.558***	1.706***	-1.438***	1.693***	0.318***	-0.041*	0.217***
Number of Satellites	-0.003	0.001	-0.004	-0.007	-0.003	-0.003	-	-	-	-	-	-
Slope (clinometer)	0.002	0.002	0.000	-	-	-	-	-	-	-	-	-
Treecover:												
Partial	0.002	0.005	0.003	0.043*	0.014	0.027*	0.054	0.016	0.044	0.016	-0.002	0.018
Heavy	0.111*	0.042	0.078**	0.080*	0.031	0.045	0.047	-0.017	0.060	0.086**	0.021	0.064**
Weather:												
Mostly Cloudy - Rainy	0.026	0.009	0.016	0.044	0.035*	0.012	-0.027	0.007	-0.019	0.018	0.017	-0.001
Psuedo-R2	0.096	0.047	0.128	0.050	0.036	0.120	0.111	0.171	0.171	0.066	0.032	0.097
N	1765	1765	1765	1908	1908	1908	486	486	486	4159	4159	4159

Standard errors clustered at enumerator level
*p<.1; ** p<.05; *** p<.01

Table 13 – Determinants of Bias (SR – CR)

OLS Regression

	Ethiopia			Nigeria		
Dependent Variable:	SR-CR	|SR-CR|	{|Bias|/CR } * 100	SR-CR	|SR-CR|	{|Bias|/CR } * 100
CR Area (acres)	0.043	0.278***	-724.752***	0.121	0.448***	-10.686***
CR Area^2	-0.247***	-	408.956***	-0.128***	0.017***	0.540***
CR Area^3	0.038***	-	-54.541***	0.004***	-	-
Number of Corners	-0.011***	0.002	-8.710***	0.047***	0.017*	1.524**
Distance from dwelling	0.012***	0.014***	4.902***	-	-	-
Number cultivated plots in HH	-0.014***	-0.011***	-3.735	0.058	0.092***	2.946
Slope (clinometer)	0.001	0.001	-1.656	-	-	-
Soil Quality (SR):						
Fair	-0.090***	-0.076***	-66.665***	-	-	-
Poor	-0.049	-0.057*	-123.089***	-	-	-
Property Rights^o	-0.012	-0.024	-33.235*	0.078	-0.129	-7.821
Treecover:						
Partial	0.131***	0.094***	117.589***	-0.270***	-0.117	-21.426**
Heavy	0.162***	0.165***	74.343**	0.518**	0.414**	-2.73
HH Head Characterisitcs:						
Female	-0.038	-0.028	-28.918	0.342	0.124	-5.362
Yrs education	0.021***	0.016***	8.871**	0.019	-0.001	-0.046
Age	-0.001	-0.001	-0.249	0.010**	0.003	0.152
Literate	-0.102***	-0.114***	-23.86	0.387	0.084	10.21
Constant	0.317***	0.231***	452.797***	-1.145**	-0.365	65.356***
N	1689	1689	1689	486	486	486
R2	0.231	0.279	0.183	0.528	0.624	0.032

*p<.1; ** p<.05; *** p<.01

^{o}Property rights defined here as: HH has title or certificate, HH has ability to sell land, *or* HH can use land as collateral.

IX. Figures

Figure 1 – Plot Size Distribution

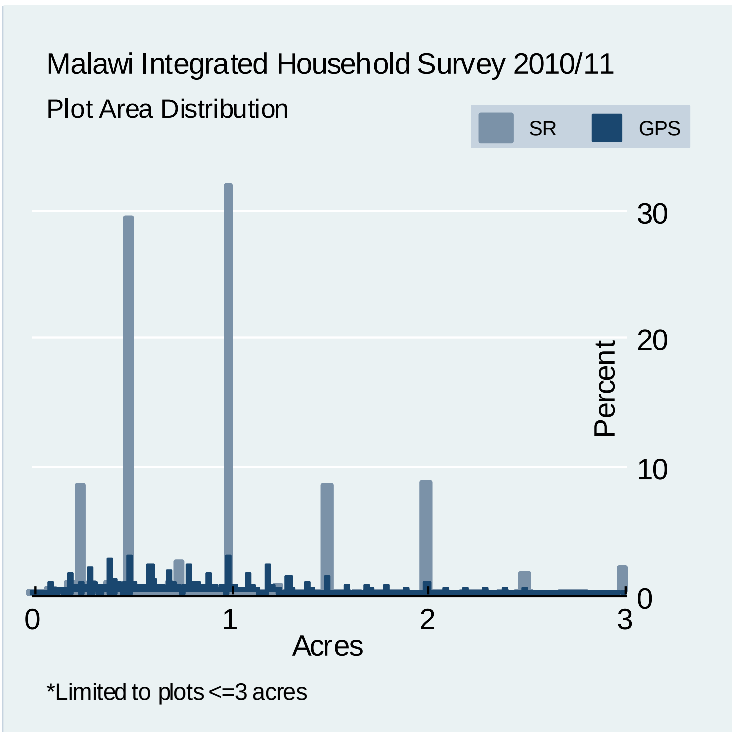


Figure 2 – Time taken for GPS and CR measurement by plot size (minutes)

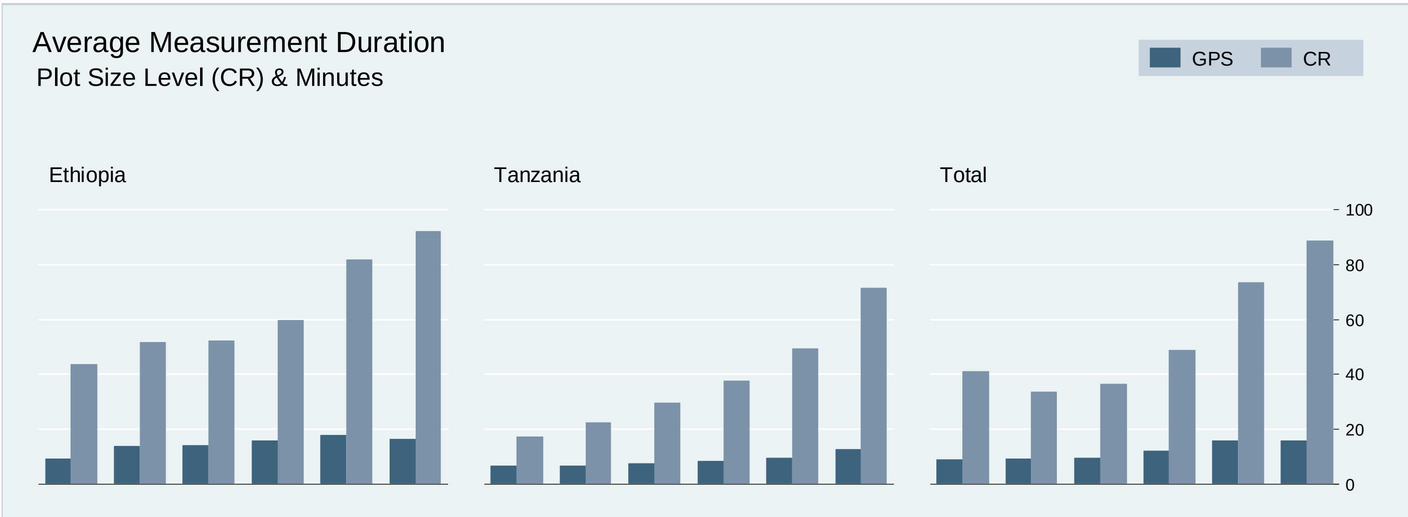


Figure 3 – Scatter plots of Compass and Rope vs GPS (left) and Self-Reported (right) land area measures, acres

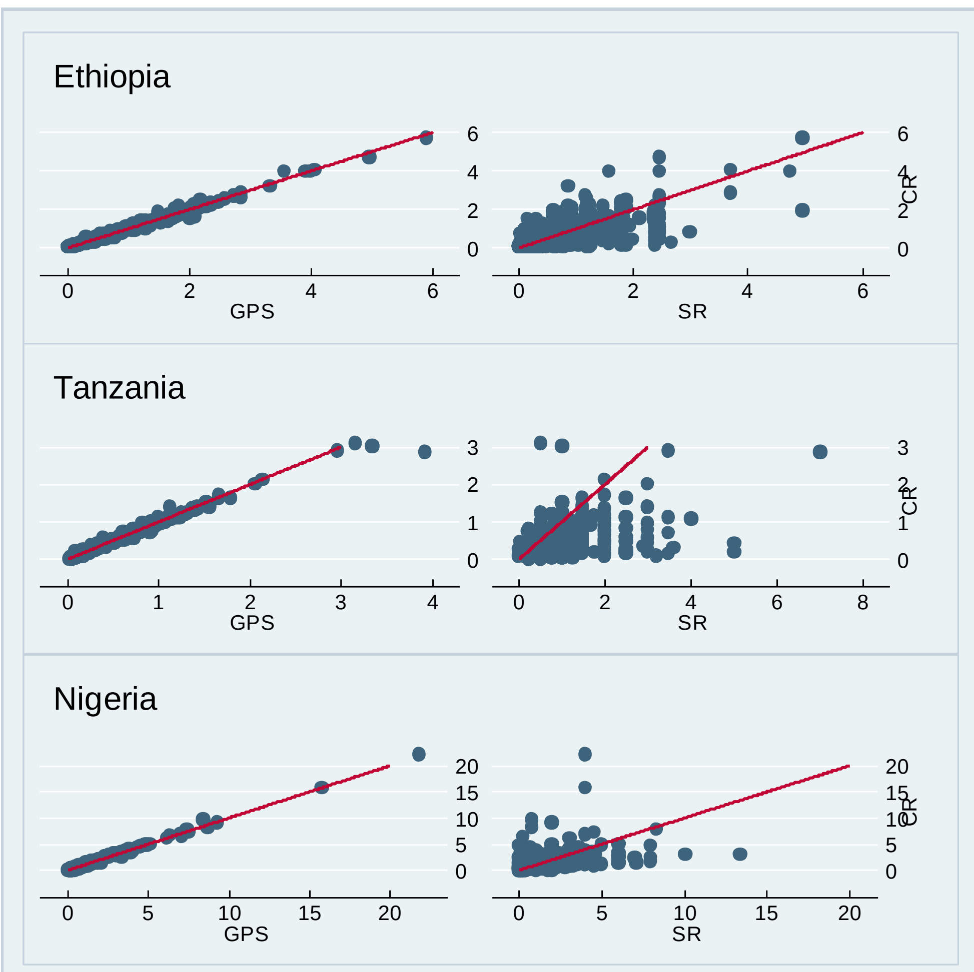


Figure 4 – Scatter plots of relative (%) and absolute (acres) bias over plot size (acres)

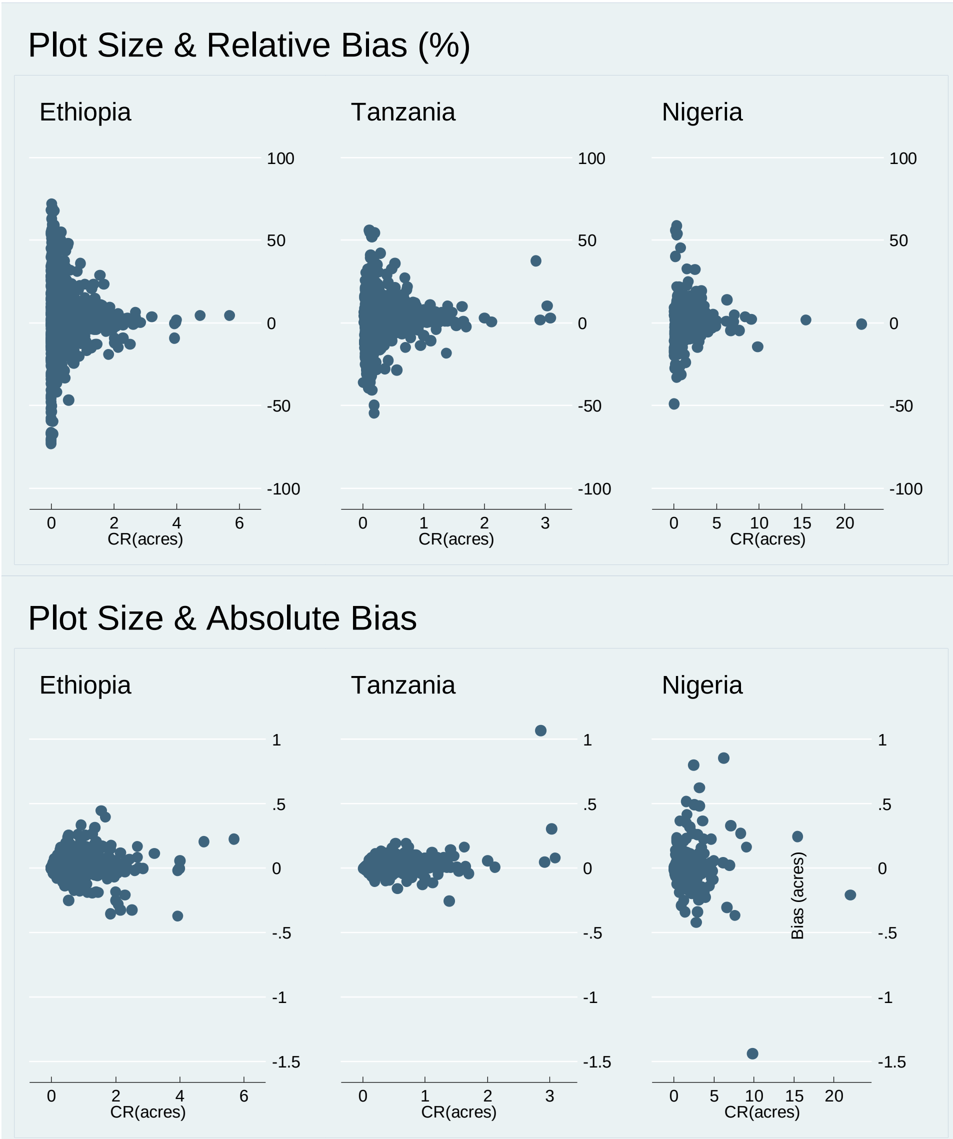
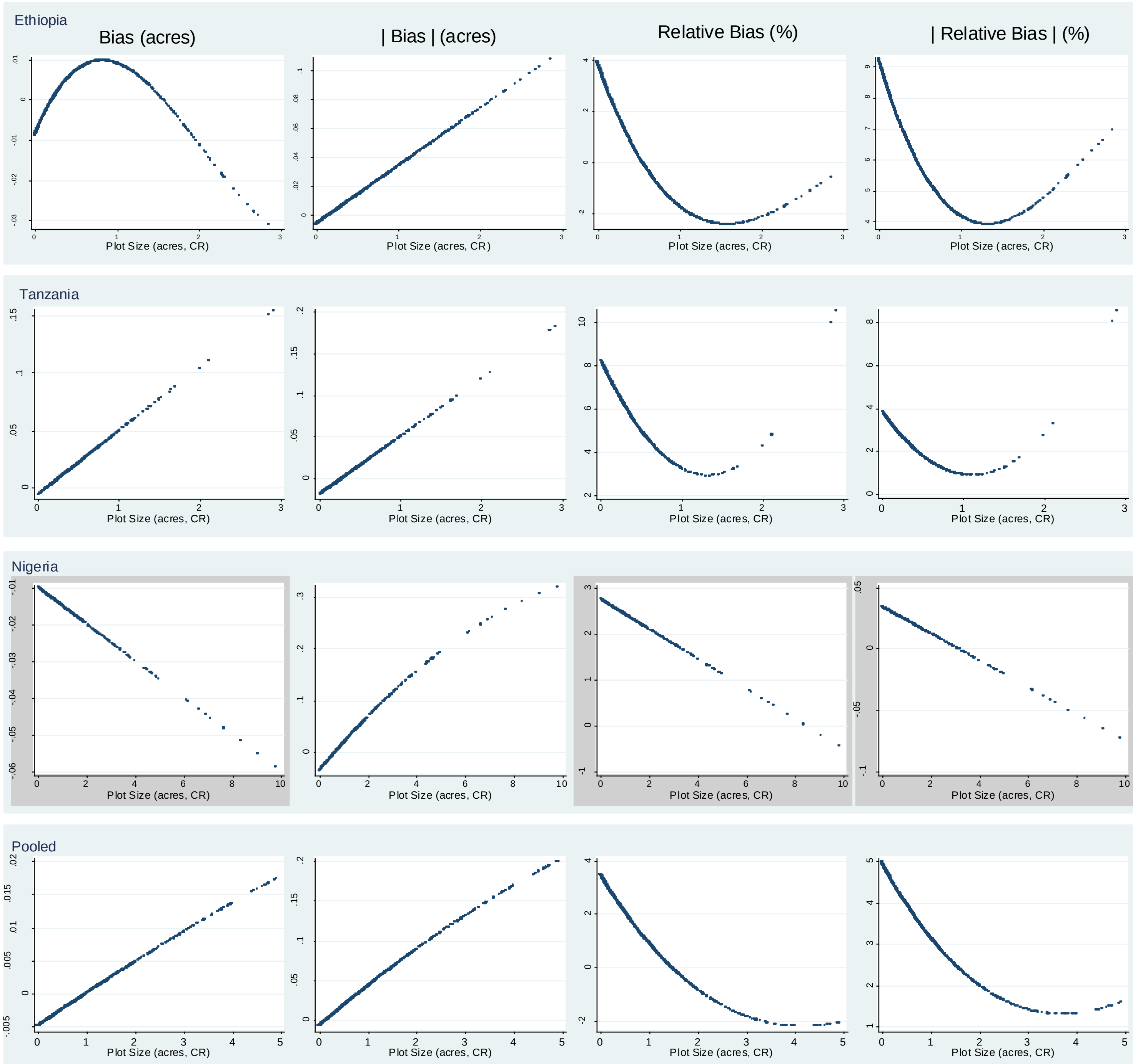


Figure 5 – Graphic representation of land area coefficients found in Table 11

Graphs consider only land area terms and constant



*Specifications in which all coefficients on land area are not statistically significant at the 10% level are highlighted in gray.

Milking the data: Measuring Milk Off-take in Extensive Livestock Systems. Experimental Evidence from Niger.

By Alberto Zezza, Giovanni Federighi[#], Kalilou Adamou[+] and Pierre Hiernaux[¥]*

**World Bank, [#]Universita' di Roma 'Tor Vergata', [+]ICRISAT , [¥]Géosciences Environnement - CNRS*

Milk is an important source of cash and nutrients for many households in developing countries. Yet, our understanding of the role of dairy production in livelihoods and nutritional outcomes is hindered by the lack of decent quality household survey data. Data on milk off-take for human consumption are difficult to collect in household surveys, introducing possibly severe biases in the computation of full household incomes and farm sales. This paper presents results from a validation exercise implemented in Niger, where alternative survey instruments based on recall methods were administered to randomly selected households, and compared to a 12-month system of physical monitoring and recording of milk production. The results of the exercise show that reasonably accurate estimates via recall methods are possible, and provide a clear ranking of questionnaire design options that can inform future survey operations.

1. Introduction and background

Despite the importance of the agricultural sector and its critical role in development policy and for poverty reduction, serious weaknesses in agricultural statistics persist throughout the developing world and are particularly pronounced in Africa. Of the 44 countries in Sub-Saharan Africa rated by the Food and Agriculture Organization, only two are considered to have high standards in data collection while standards in 21 countries remain low (Carletto, 2009).

Statistics on livestock stand out as an area in particular need for improvement. There are important technical reasons, besides institutional and political neglect, that explain why livestock data are particularly scarce or of dubious quality. Unlike crops, which are rooted in a specific tract of land and can be counted and measured, livestock are mobile, posing a challenge to enumeration even in sedentary livestock systems. The difficulties of collecting data on livestock are exacerbated by peculiarities in the management of livestock assets, in the mobility of some population groups that are especially reliant on livestock for their livelihoods (e.g. pastoralists), and by the fact that livestock products tend not to have one or two specific harvests at predetermined points in time, but tend to be produced either continuously or irregularly throughout the year, often with seasonal patterns.

The need of addressing the current shortcomings in the quality and availability of livestock statistics is only made urgent by the rapidly increasing importance of the livestock sector. In developing countries as a whole, milk consumption almost doubled, meat consumption tripled, and egg consumption increased by a factor of five in the past fifty years, in what has been dubbed a ‘livestock revolution’ (Delgado et al., 1999). During the same period consumption of cereals increased only slightly and that of root and tubers actually declined (Gerosa and Skoet, 2013).

It is not uncommon for shares in excess of 60 to 70 percent of rural households in African countries to hold some livestock and depend on it to some extent for generating income or accessing nutrient dense foods (Davis et al., 2010). In rural Niger 3 out of 4 households keep some livestock according to recent national household level data (Bocoum et al., 2013), and according to FAO Statistics for 2010 (FAOSTAT, 2015) livestock contributes 28 percent of the net value of agricultural production in the African continent, with values substantially above average in several countries. In West Africa the share of livestock in agricultural production is consistently large, being 43 percent in Mali, 30 percent in Burkina Faso, and 38 percent in Niger, but other major countries also record sizeable shares, such as Ethiopia (35 percent).

Part of the neglect of livestock statistics materializes in underinvestment in actual data collection, but even when data are collected their quality is uncertain because of a lack of rigorous methodological work to assess the reliability of data collection practices. This paper aims to contribute to improving the practices for data collection on one specific item, milk, of major importance for livelihoods, income, food security and nutrition in many parts of Africa based on fieldwork in one of the African countries, Niger, where livestock constitutes the backbone of the rural economy.

Without reliable and timely livestock statistics it is hard to see how countries such as Niger can design, monitor and evaluate effective policies for promoting the role of livestock for poverty reduction and food security. The lack of high quality data on the dairy sector hinders both advocacy and policy analysis efforts aimed at informing actions to support livestock-based livelihoods. Household-level data and studies on the role of milk off-take for human nutrition and livelihoods are severely hampered by the difficulty of producing reliable estimates of milk off-take in small-scale livestock production systems.

Milk production offers an important source of cash income to many of the over 200 million poor livestock keepers estimated to reside in developing regions (Thornton et al., 2002; Pica-Ciamarra et al., 2011). For pastoral communities milk is often the sole source of calories and key nutrients, and a major source of cash income (Sadler et al, 2009). Some livestock products such as milk and eggs can help poorer households mitigate the effects of often large seasonal fluctuations in grain availability (Wilson et al., 2005). Hoddinott et al. (2014) using Ethiopian data found empirical evidence to support the hypothesis that cow ownership in underdeveloped rural settings is a key driver of the milk consumption and linear growth of young children.

From a nutritional point of view, milk is a good source of dietary fat, energy, protein and other nutrients (Wijesinha-Bettoni and Burlingame, 2013) that brings “important nutritional benefits to large segments of the population of developing countries” (Muehlhoff et al., 2013: p. 5). In particular, milk can provide substantial amounts of nutrients such as calcium, magnesium, selenium, zinc, riboflavin, vitamin B12 and pantothenic acid (Weaver et al., 2013). Milk can help provide children of age 6-24 months that are not being breastfed adequate quantities of fat, which is crucial in their diets because it contains essential fatty acids, facilitates the absorption of fat soluble vitamins, and enhances dietary energy density and sensory qualities (Dewey, 2005).

Milk consumption has also been associated with secular growth in height whether in industrialized and developing countries (Japan, India) or in pastoral societies (Weaver et al., 2013; Hoppe et al., 2006). A review of the available evidence, laments that despite the observed increase in milk production and consumption world-wide, child undernutrition and micronutrient deficiencies that could be alleviated by increased intake of milk and other animal source foods remain highly prevalent. In developing countries, both milk and meat intake improve growth indicators, micronutrient status, and cognitive performance (Dror and Allen, 2011).

In general, it is hard to appreciate the role of milk and dairy production in household level livelihood studies in developing regions, because of the generally poor state of agricultural statistics in these countries, and because of the practical difficulties in measuring milk off-take in household surveys. Milk off-take is difficult to measure in household surveys because: (a) Lactating females can be milked daily (often twice, mornings and evenings), but with seasonal patterns; (b) Milk varies depending on the lactation stage; (c) Milk can be left to feed young sucklings; (d) Reproductive and lactating females may be present but not necessarily being milked. These potential sources of measurement error combined make the valuation of milk off-take particularly challenging in household surveys, introducing possibly severe biases in the computation of full household incomes and farm sales.

This paper presents results from a validation exercise implemented in Niger, where two alternative survey instruments were administered to randomly selected households, and then compared with the results of a physical monitoring of milk off-take over a 12-month period. The immediate objective of this work is to draw lessons for questionnaire design by selecting the best performing options and identifying outstanding issues. The ultimate goal is to contribute to a better understanding of the role of animal production in livelihoods and nutrition, which can facilitate more effective policy and program design.

The focus in the paper is on one specific family of household surveys, the Living Standard Measurement Study (LSMS). This is one prominent type of household survey widely implemented in developing countries to monitor and analyze poverty and livelihoods. While this is just one example of a multi-topic household survey for livelihood analysis, we maintain the lesson for questionnaire design assessed with this exercise can be applied beyond LSMS surveys. The paper is organized as follows. The next section outlines the overall design of the validation exercise and the survey instruments being tested. Section 3 describes the data, and section 4 presents the results. The concluding section discusses the implications of this work for future data collection, and elaborates on ongoing next steps in furthering this line of work.

2. Testing alternative survey instruments

2.1 The context: Survey validation work in developing countries

In their primer on methods for testing and evaluating survey questions, Presser et al. (2004a, p: 109) note how "pretesting's universally acknowledged importance has been honored more in the breach than in the practice". Even in countries with well-oiled and well-financed statistical systems, pretests are often limited to a rehearsal of survey interviews, usually on a fairly limited number of cases, which are then qualitatively evaluated by the survey teams so as to draw lessons on questions that seemed to pose problems to interviewers or respondents. Sometimes, this is complemented by a quantitative analysis of response frequencies and other simple statistics from the data collected during a pilot survey.

In most cases there is little that is systematic about these tests, despite the existence of techniques geared towards assessing the performance of survey instruments (see e.g. those reviewed in Presser et al., 2004b, and Iarossi, 2006), and very little documentation is provided to users of the data on the contents of such tests. The evaluation of what 'works' is mostly left to the judgment and experience of the survey team.

Increasingly, however, survey practitioners are paying attention to pre-tests as means towards improving data quality. Also, specific methods are being developed, tested and codified and increasingly applied in survey practice. The interested reader is referred to Presser et al. (2004b) for an excellent review of methods such as cognitive interviews, behavior coding, response latency, vignette analysis, experiments, and statistical modeling. While the use of such methods, and their documentation, is more commonly found in OECD country surveys, their application is gaining grounds in low income countries, including in Africa.

Despite the fact that the quality of the data should be of interest to researchers as much as the quantity, it is surprising how little attention the formal validation of household survey data collection has received in the literature. Researchers' preoccupation with data quality results mostly in efforts to design and supervise survey work as well as possible, but very infrequently are the results of such efforts formally tested. There are some notable exceptions however, and our study aims to contribute to this small but growing strand of methodological literature.

Most of the existing literature on survey experiments and survey validation refers to the measurement of household consumption. Beegle et al. (2012) test eight alternative methods of measuring household expenditure, comparing personal diary as the benchmark to other diary and recall formats. They find significant differences between resulting consumption measures, with the correlation between under-reporting and both illiteracy and urban households' status being particularly evident. In addition, Gibson et al. (2013) use data from the same survey experiment in Tanzania to obtain evidence on the nature of measurement errors, concluding that, as expected, errors have a negative correlation with the true value of consumption.

In the context of household consumption, another issue that has been analyzed is the extent to which the length of the lists of consumption items affects estimates of household expenditures. In a study in El Salvador, Joliffe (2001) shows that a more detailed consumption list, results in higher estimates of mean household expenditures (by around 30 percent). This finding has clear implications for the resulting poverty estimates.

The impact of the level of detail of the questionnaire on key indicators has also been investigated in the field of labor market statistics. Dillon et al. (2012) consider if this aspect, together with the type of respondent, can explain the existing widespread variation in measurement of child labor statistics.

Scott and Amenuvegbe (1990) conduct an experimental study on 135 households in Ghana. Each of them was interviewed 11 times at varying time intervals, asking to report expenditure on the 13 most frequently purchased items. In this study, each additional day of recall returns in a 3% decline of the reported daily expenditure.

The choice of reference period is also likely to have considerable impact in several domains. Beegle et al. (2011) test for recall bias in agricultural data, submitting questionnaires with different length of time between harvest and interviews for three African countries. An assessment of whether and how modalities of data collection in agricultural production may affect results is also provided by Deininger et al. (2012).

Using data from two microenterprise surveys in Sri Lanka, De Mel et al. (2009) find that firms under-reported revenues by about 30%, and that requesting them to maintain account books had significant impacts of on both the revenues and expenses they reported, but not on profits. More generally, they argue that questions on profits give truer measures than asking about revenues and expenses.

What this literature shows is how data collection methods matter as much as analytical tools and statistical techniques for the conclusions of a study. Yet, researchers are often ill equipped for judging the extent to which data quality can be affecting their results, whether using data collected by others or data collected as part of their own research, as the survey instruments employed rarely undergo this type of systematic validation. In particular, we are not aware of similar work done for livestock questionnaire design in the context of household surveys in low income countries, which is the reason that motivated a joint effort by FAO, ILRI, and the World Bank (as part of the LDIA and LSMS-ISA projects[1]) to start the survey validation work that is documented in this paper.

2.2 *Milk off-take recall methods*

LSMS surveys have typically lumped the collection of data on livestock products in one table listing the different products on the rows and a set of standard questions, common to all products and based on a 12-month recall period, on the columns. The module usually asks a variation on two rather simple questions: (1) “Number of production months in the last 12 months”, and (2) “Average production (off-take) per

[1] For information on the two projects see http://www.fao.org/ag/againfo/programmes/en/Livestock_Data_Innovation_in_Africa.html and www.worldbank.org/lsms.

month during production months". Sometimes these questions are asked for milk as a homogeneous product, sometimes the product is broken down by livestock species (cow, ewe, goat) or by dairy product (fresh or curd milk, cheese, butter).

Because of the peculiarities of milk production recalled earlier (continuous production, seasonality, varying lactating capacity of animals, including over time), such simple recall questions are likely subject to large errors. This has led livestock researchers and livestock survey specialists to devise more complex strategies to collect more accurate milk off-take data as well as an expanding set of additional information useful to evaluate milk production systems.

Examples of these elaborate approaches include the 12-month method developed by researchers in France's CIRAD (see Lesnoff et al., 2010), which relies on the monitoring/recording of off-take over extended periods of time, as well as on techniques which while based on recall approaches, try and prompt the respondent more in depth about the milk off-take system hoping that this will help increase the accuracy of the responses. In developing new survey approaches to be integrated in LSMS-type surveys that include an expanded agricultural focus, these approaches are useful, but need to be adapted to conform to both the objective of the survey as well as to the survey operations. The only way to assess whether a change in approach results in an actual improvement in data quality is to validate the new method via fieldwork, ideally in an experimental setting, while reproducing as closely as possible real survey conditions.

The main goal of LSMS-type surveys is to generate information on household living standards and livelihoods, in this case jointly with information on the productivity, profitability and returns to different activities households may be engaged in. The LSMS survey logistics are organized with mobile teams, that normally reside in each enumeration area for 3-4 days, and need to complete the survey operations in that location in that given time. It is therefore beyond the scope of the LSMS, in terms of both objective and logistics, the possibility to collect milk off-take data over extensive time periods, or in a way that allows calculating the complex milk productivity parameters often required by livestock sector specialists. The objective of an LSMS needs to be more modest, and limited to collecting a reliable measure of milk off-take that can accurately portray the role that milk has in the overall household livelihood strategy.

At the same time, LSMS-type surveys aim to look at the heterogeneity across households, so methods that rely on the application of technical production factors from the literature (e.g. average milk production or off-take per animal in a certain environment) to variables that may be easier to measure in a survey (such a number of animals milked by the household) may result in accurate 'average' estimates, but may artificially reduce the observed differences in milk off-take (both in physical and value terms) across households. For most of the analysis performed with LSMS data, the analysis of the dispersion of the distribution is often as if not more important than the analysis of the measures of central tendency (means, medians). Also, the number of lactating cows, ewes, goats milked, the volume of milk extracted, the amount of time milking is practiced for are all management decisions that vary across households and herders, for reasons that include but go beyond the milk production potential of the animal as expressed by technical parameters. For these reasons, competing data collection methods will need to be evaluated not only on the basis of their ability to yield an accurate point estimate of, say, mean milk off-take, but also on their ability to return a distribution of observations that resembles as much as possible the 'true' distribution .

In view of these considerations, in developing the Niger survey validation we looked at two methods that are often applied in livestock sector surveys, but also seemed to hold promise of being adaptable to both the questionnaire design and logistics of LSMS survey operations. In what follows we will refer to these two methods as the "Average Milk per Day" (AMD) and the "Lactation Curve" (LC) methods.

The two questionnaires are identical, except for one question on milk off-take. Both questionnaires are asked at the level of each animal species (cows, ewes, goats, camels), and start off by prompting the respondents about the number of months during which animals were milked for human consumption, and how many animals were milked on average during each of those months. The questionnaires then differ in that the AMD asks for the average quantity per day off-taken during the reference period[2], whereas the LC questionnaire asks about the amount of milk off-take on average from the animals milked at three, or four, different points in time: one week, one month, three and six months after parturition. The two modules then continue asking the same set of questions

[2] In fact, there are two variants of the AMD method, one that asks about Average Milk *per Animal* per Day (AMAD), and one that asks about Average Milk *per Herd* per Day (AMHD). Our results refer mainly to the latter even though – as we explain in what follows – we used both at different points in our fieldwork.

on issues of whether calves/lambs/kids were allowed to suckle, about the time duration between parturitions, and about the disposition of milk off-take (consumption and sales either fresh or after transformation into dairy products).

Annual household milk off-take can be calculated from both questionnaires. In the AMD this is done by simply multiplying the average daily off-take by 30 days (to get to monthly off-take per animal), then by the number of off-take months and by the number of animals milked. In the LC methods things are a little more complicated, and annual off-take is calculated as the area under each animal's lactation curve, or rather milk off-take curve. This is not immediately intuitive and requires some further explanation.

All mammals have a lactation pattern with lactation starting shortly after parturition, a peak reached early in the lactation period, followed by a slow decline to the end of the lactation period. The timing of these periods, and the overall length of the lactation vary by animal species, and by breeds, and with climatic, grazing, watering and a host of other factors. Besides that, what the survey measures is not lactation as such, but the amount of milk that is taken off for human consumption, which is a decision variable for the farmer.

Total milk off-take can therefore be approximated, assuming a constant value of off-take between the last point in time for which recall is asked and that of the end of the milking period[3], as the area under a curve such as the one depicted in Figure 1. In the most general case of four monitoring points, the corresponding formula can be written as:

$$Q = q1m*30+(qs-q1m)*30*0.5+q3m*60+(q1m-q3m)*60*0.5+q6m*90+(q3m-q6m)*90*0.5+q6m*(end-6)*30$$

Where *Q* is the total milk off-take per animal in one lactation, *qs* is the average daily quantity of milk off-taken per animal at the start of the lactating period (one week after parturition in the Niger LC module), *q1m, q3m* and *q6m* are respectively the off-take one month, three and six months after parturition, and *end* is the average number of months

[3] An alternative way of computing milk production is to assume a monotonic decline in lactation from a peak after a week from parturition, to zero at the end of the lactation (milk production) period.

of milk off-take per animal. For animals with shorter lactation periods such as ewes, and goats, more parturitions (and hence lactating periods) may fall within the 12 months of the survey reference period. In such cases, the presence of a question on the average interval between parturitions allows attributing a quota of the second lactation to the survey reference period (Njuki et al., 2011). In this paper we focus on cattle milk over a 12 months reference period, which rules out the possibility of multiple lactations for any one animal as the calving interval for cattle is longer than 12 months.

With the LC method respondents are asked to recall more information (milk off-take at different stages of lactation) but to only average out this information across the animals they have milked. In the AMD method, respondents are required to report only one figure, but to obtain that via an implicit process of averaging not only across animals but also across lactation stages. What process is easier for the respondent and more likely to return an estimate closer to the 'true' value is an empirical question, and the main question this paper aims to address. Whether it is easier for respondents to respond to questions about an average animal or about the entire herd is also an empirical question. In the study area each animal is milked separately but the milk extracted is poured in a single pot (or a series of pots), thus the herder in charge of milking may have a feeling for both the average volume of milk from a cow and the average volume of milk collected from the herd. After some piloting in the field, it was felt that respondents found it easier to report about off-take per herd, as the milk is collected for all animals into one container, once or twice a day.

It should be noted that in empirical applications, particularly in specialized dairy livestock surveys, the LC methods is often implemented with respect to one or more specific animal(s) selected at random from the respondent's herd. This also allows capturing the possible variation in the lactation stage of different animals throughout the year, where this is a concern. In living standard surveys, where livestock is just one component of a more complex survey, this would not however be practical. Interviews are often carried out at the household residence, and the herd may not be physically present for the selection of the animals to be made. Also, if both large and small ruminants need to be enumerated that would require identifying different animals in potentially different locations at the time of the interview. The time and cost implications of this for a large scale national survey are likely to be prohibitive. The results presented below therefore need to be interpreted with the qualification that they apply to the lactation curve method as applied in this exercise. Particularly when there is variability across animals and over time, respondents may find it easier to recall milk off-take at specific points of the lactation for specific animals as opposed to an hypothetical 'average animal' as implied by the way the method was applied in the Niger survey we report about.

In the study area milk off-take has a markedly seasonal pattern, and animals tend to follow similar patterns over the year so that one does not expect to find significant differences in the lactation stage across animals at any point in time. The seasonal increase in number of cows milked and the increase in volume of milk off-take by cow, both contribute to the increase in milk off-take by farm during the wet season (August to December), then progressively decreasing reaching the minimum off-take level from March to June. Variability in the lactation stage across animals is therefore not a major concern in the area the data collection for this study took place.

FIGURE 1 HERE

Some livestock survey practitioners suggest that the response given to the AMD question may result in an overestimate of the quantity of milk collected as the response patterns may lead to estimating the area under a rectangle that will largely be above the lactation curve triangle. Figure 2 illustrates the point, using hypothetical values not too dissimilar from the data in our Niger cattle milk off-take study. In calculating total milk off-take from the AMD method one is essentially computing the area of the rectangle ABCD, where AB is the number of months milk was collected and BC is the monthly quantity (in liters) collected milk[4]. Suppose the true shape of the off-take curve for the respondent was equal to the line BEF, and it becomes evident how AMD would result in an overestimate of milk off-take.

The AMD method can be administered for different recall periods, as it is often argued that shorter recall can improve data quality. This is especially true for variables that are characterized by seasonal patterns, which is the case for milk off-take. In the case of the LC method this is not feasible as a, say, 6-month recall period would likely be shorter

[4] Note that the area of the rectangle ABCD depends on the mean milk off-take by animal by day (BC) but also of the mean duration of milking (AB). The latter is much more variable and depends on factors such as individual, parity, season, and it could reflect a progressive reduction of milking frequency.

than the lactation/off-take period, thus complicating the task of the respondent as some of the points in the off-take curve may fall outside of the recall period.

FIGURE 2 HERE

It is also often found that additional questions related to the main object of the recall can be useful in aiding the recall by the respondent. For that reason, in the exercise described in this paper, we also experimented in combining the AMD with the LC questions. The idea is that if a respondent may provide a more accurate answer when asked to estimate average off-take if she is also invited to recall average off-take at different stages over the lactating period, than if asked to provide that figure directly.

In the exercise reported on here we compare the following methods: (a) the LC method; (b) the AMHD method with a 12-month recall; (c) the AMHD method with a 12-month recall and linked to the LC method questions; (d) the AMHD method with a 6-month recall. All are compared against a benchmark constructed by the physical monitoring of daily milk off-take measured every fortnight over a 12-month period. We also provide some evidence on the performance of the AMAD variant of the AMD method. Before discussing the results of these comparisons, we now turn to a description of the data.

3. Data

The main data set analyzed in this paper comes from fieldwork that took place in the Dantiandou district in Niger, between April 2012 and June 2013, and is referred to here as the Dantlait survey. The fieldwork was managed by two experienced enumerators, and a supervisor, all three ICRISAT staff. The team monitored the milk off-take of 300 families over 12 months, as well as associated livestock management, together with family consumption and sale of dairy products. The team also administered 6-month recall questionnaires on 200 families, and 12-month recall questionnaires to 400 families.

The first 200 family farms were randomly sampled among the 835 family farms documented in 2009 and 2010 for the Livestock Climate and Society (ECliS) project[5].

[5] Final report and documents available at http://eclis.get.obs-mip.fr/ .

These 835 families live in 13 villages and associated camps within the district (commune and canton) of Dantiandou (80 km East of the capital Niamey). A large data base is available on the composition of the family, its economic activities (including cropping, breeding livestock, forestry, and off farm), the composition and number of livestock, milking practice, consumption and sale of dairy products. This data base was used to stratify the families based on the type of dwelling (either village or camp), which largely matches with socio-ethnic affiliation (Zarma/Fulani), and on the size of the herd (less than 5, 5 to 15, more than 15 adult females). The additional 100 families were selected in 13 additional villages from the district of Dantiandou (5,340 families and 45 villages in total), based on the 2008 national census.

The monitoring method targeted the assessment of the daily milk off-take in each of the 300 sampled families. For each sampled family herd, the milk off-take was measured one day every fortnight adding morning and evening milking when applicable. At each milking, the total milk off-take of the herd was poured in a transparent plastic pot devoted to that measure. The level reached by the milk was marked on the outside of the pot with a marker by the herder. To assess milk volume, the research assistant weighted the plastic pots empty and when filled with water up to the mark done on the side of the pot. The pot weights were recorded on the herd recording form together with the number of lactating females, and the number of lactating females milked re-actualized at each visit. Equipped with a motorbike, each of the two enumerators monitored about ten farms per day (one or two visits depending on milking practices), with revisits every two weeks.

Camp families involved in dairy farming are endowed with larger cattle herd on average than village families (7.2 vs 4.4), both however are managing quite small herds. The mean number of lactating cows in the course of the year is 3.4 vs 1.8. Only a fraction of the lactating females are actually milked, on average 1.9 vs 1.3. Resulting mean milk off-take is low, at 2.1 liters per day in camps and 1.3 liters per day in village farms. There are large seasonal variations, the wet season and first part of the dry season (‘cool’ season) contrasting with the late dry season, with milk yield in a factor 2 in camp farms and factor 1.5 in village farms. These seasonal variations are explained by the reproductive cycle of the cows (peak of birth in early wet season), the better quality of grazing resources, but these reasons are mediated by the herder decision (i.e. share of the lactating cows actually milked, milking in the morning/evening or both, volume off-take). It appears for example that the volume milked (0.8 to 0.9 liters per cow and per milking) does not vary with farm type, morning or evening milking, position along the lactating curve. Sparing milk for the calves drives the practice of milk off-take especially in camp farms.

Recall questionnaires were asked to 200 farms (141 of which had also been monitored) in December 2012, and to 400 farms (269 of which had also been monitored) in May-June 2013. The December survey included a 6-month recall AMHD questionnaire. The 400 households interviewed for the 2013 survey were randomly split into two groups, with one being administered a 12-month AMHD recall, and the other a LC module, where the LC questions followed an AMHD question. We are therefore able to compute recall measures based on the four measures described above (6-month recall, which we also annualize by multiplying it by 2), LC curve, 12-month AMHD, and 12-month AMHD cum LC recall aid.

The objective of the physical monitoring was to construct a measure that could be used as a benchmark against which the different recall methods are compared[6]. Earlier in the project, a LC questionnaire and a 12-month AMAD recall had been included in the national ECVMA survey implemented in 2011 by the 'Institut National de la Statistique' (INS) of Niger, with technical assistance from the World Bank and the 'Ministère de l'Elevage', on a nationally representative sample of 3,968 households, of which 2,430 are rural and 1,538 urban. While it was not possible to construct a benchmark for this large nationally representative survey, the results of the comparison between the two recall methods can be interpreted in the light of the conclusions emerging from the Dantiandou survey and monitoring.

While the standard LSMS-type livestock product module was not used in these surveys (national ECVMA and Dantiandou monitoring surveys), a smaller scale pilot survey that was run in February 2011 on about 60 households provides qualitative confirmation of the expectation that the standard LSMS module tends to understate milk off-take compared to other recall methods. As the ECVMA national sample, the pilot survey

[6] One source of measurement error we do not focus on here, is the tendency of respondents to recall question to report figures in round numbers (e.g. quarter of a litre or half a litre), whereas for physical monitoring measures were recorded in actual cubic centimetres. This is a well know phenomenon known (see Roberts and Brewer, 2001). For an application to land size measures in agricultural surveys see Carletto et al., 2013.

asked the milk off-take quantities not only for the entire herd but also for each different livestock species (ewes, goats, cows and camels).

Of importance to the design of the study, we observe no significant differences between the two groups in which our sample was randomly split. That provides confidence in that the random design on which the survey is based worked, and that the groups being compared have no systematic difference other than the fact that they have been asked different questions. Table 1 summarizes the descriptive statistics for the key groups in which the sample has been split for the fieldwork and the analysis. On only one variable (months of milking) we do find statistically significant differences between the LC and the AMD questionnaires. However, for a further separation of those who responded to the LC or to the AMD questionnaires into the monitored and not monitored groups, there are no statistically significant differences. Most of the comparisons we will base our conclusions on will bear on the 269 monitored households only, so that even if there was a bias in the selection of the households to monitor, it would not affect the comparisons. The non-monitored households were mainly added to the sample to obtain some more statistical power in the comparison of means.

TABLE 1 HERE

4. Results

The expectation going into the exercise was that the LC method could provide an improvement over the AMD, which we expected to overestimate off-take. The key results from the validation exercise carried out as part of the ICRISAT-led fieldwork in Dantiandou are reported in Table 2.

The first rather surprising result is that the AMD recall methods do in fact perform much better than was expected, and appear to be superior to the LC methods. The deviation of the median values from the median of the milk monitoring is surprisingly close to the value obtained via the physical monitoring with a difference of just 21 liters (or about 3 percent). The deviations for the mean values are somewhat larger but still acceptable at 30 liters (3 percent of the monitoring value for the 6-month recall, up to 6 percent for the other variants).

Secondly, for the LC Method the results are less satisfactory. Deviations from the 'gold standard' represented by the physical monitoring range between 'acceptable' levels at 6 and 10 percent, when median values are considered (for the 4- and 3- point measures, respectively). If one considers deviations from the average value of the monitoring, however, differences increase to 13 percent for the 4-point LC method and 37 percent for the 3-point LC method. In general, the 4-point method appears to perform significantly better, thus justifying the extra question required of the respondent.

Thirdly, the results show that a major feature common to both the AMD and LC methods is how they over-estimate the dispersion around the mean (as measured by the standard deviation), and particularly so for the LC method. Among the AMD variants, the highest standard deviation is 1.4 times the standard deviation of the monitoring. For the LC method the ratio is 1.8.

Within the AMD methods, shortening the recall period to 6 months appears to perform as well as the 12-month recall, without any major improvement in accuracy. In this particular sample the 6-month recall survey did not generate any large extreme value, which happened for the 12-month survey, but it is hard to generalize this result, as it is linked to the performance of a few respondents[7]. We will however return to this matter when we discuss other measures we use to assess the relative accuracy of the various methods.

Another result that is interesting to note is how the AMD method, when integrated with the LC questions, returned substantially more accurate results than when the LC questions were not included. It appears, but again this is based on few observations in the left hand tail of the distribution, that introducing LC as a recall aid did help respondents to average out to values closer to the 'true' value, which we approximate here with the monitoring. This is particularly true for camp household, which are characterized by both higher off-take values, and deeper seasonal fluctuations.

Importantly, very similar findings regarding the differences between the estimates obtained via the AMD and LC methods are observed in the data collected via the national ECVMA survey, which did not include a benchmark measure as did the Dantlait

[7] One possible reason is that the 6-month recall survey occurred at the end of the wet season, so that the average milk offtake was mostly based on the high milk off-take during wet season while the 12-month survey occurred at the end of the dry season, requiring for the herder a more difficult averaging exercise between low milk off-take in the last months and higher off-take in the former months.

survey. Figure 3 reports the mean, median and standard deviation measures of milk off-take per cow in both data sets. The patterns, in terms of differences between the LC (based on 3 points) and AMD methods, are very similar in the two surveys. This is consistent with the idea that the Dantlait survey results can be extrapolated to a sample of households in other parts of Niger, conducted as part of a larger scale, national survey operation conducted by the national statistical office.

FIGURE 3 HERE

It is important to note, however, that the Dantlait survey was limited to cattle. Small ruminants have shorter lactating periods, and the same results may very well not apply to them. In the ECVMA data, for instance, milk off-take from ewes and goats[8] is substantially higher when estimated with the AMD method compared to the 3-point lactation curve method, which is the opposite of what happens for cattle in the same sample. Unfortunately, as discussed earlier, the ECVMA did not include a benchmark that allows assessing the precision of these estimates, and the Dantlait survey only collected data on cattle. Throughout the paper, therefore, we will be referring only to estimates of cattle milk off-take.

Besides getting at reasonable average estimates, however, LSMS-type surveys are geared towards depicting the heterogeneity in household's livelihoods and productivity. To that end, looking at how different indicators perform along the entire distribution, and understanding how well they can estimate the position of each household along the distribution is as, if not more, important as obtaining an accurate central tendency measure. For these reasons it is worth analyzing also the correlation and regression coefficients between the different recall methods and the monitoring benchmark (Table 3), and the box plots for the different measures (Figure 4).

Looking at Table 3, it is comforting to observe that the implicit ranking of the different recall methods observed for central tendency (Table 2) is also confirmed when one looks at the overall correlation between the measures resulting from different recall methods. The annualized 6-month (AMD, top row) and the straight 6-month recall (bottom row)[9] display the highest coefficients and R2, followed by the other 12-month recall methods

---

[8] Data not reported, but available from the authors upon request.
[9] The annualized 6-month recall is just the 6-month recall times 2. What really changes in the comparison between the two is the benchmark data, which is the full 12 months of monitoring in the former case, and the first 6 months of monitoring in the latter.

in the order in which they appear in the table, and again pointing to a better performance of the 4-point compared to the 3-point LC variant.

TABLE 2 HERE

TABLE 3 HERE

The box-plots (Figure 4) provide further support to these results. To improve readability we have only graphed five indicators, the monitoring, 6-month recall annualized, 12-month recall, and 3- and 4-point lactation curve. As in the statistics shown in tables 2 and 3, the 6-month recall method shows a little more dispersion than the monitoring, but in terms of median and overall distribution the fit is overall very good. The dispersion at the top end of the distribution increases with the less precise AMD methods, but remains broadly acceptable (even though it is of course hard and to some extent subjective to define 'acceptable' in this case), and becomes substantially higher for both variants of the lactation curve method.

FIGURE 4 HERE

Then, to look more closely into the correspondence between the different measures for the individual households, we have plotted scatter plots of the different recall measures against the result of the milk monitoring. Results are reported in Figure 5, where the green line represent the line of equality between the two measures (this would be a 45 degree line if the axes had the same scale), whereas the blue line is based on a linear fit.

FIGURE 5 HERE

A few things are notable from these graphs. First, the methods that perform better when judged on the synthetic measures we have analyzed so far, also perform better when we look at individual household observations. The cloud is a lot more scattered in the case of the LC method than it is for the 6-month recall or the 12-month recall with the LC aid. Second, a fair amount of measurement error remains10. More importantly, at this visual inspection the error does not seem to be randomly distributed, but tends to be negatively correlated to 'actual' (i.e. monitored) milk off-take. Respondents are more likely to under-report milk off-take if they produce larger quantities of milk, and they are more

[10] It should be noted that while we treat milk monitoring as the benchmark this measure is also, as any measure, affected
by some degree of error.

likely to over-report off-take when they produce smaller quantities. This is clearly a matter of concern for the analyst, as measures of income from milk off-take and productivity based on such data would be biased on ways that are correlated with other variables of interest.

For that reason, it is important to understand what are the correlates and determinants of the observed measurement error. Table 4 presents the results of a series of linear regressions where the percentage difference between the recall methods and the monitored milk quantities (the dependent variables, one for each method) are regressed against a set of covariates which we expect to be able to influence the quality of the recall. The independent variables in the regressions include herd and production system characteristics, as well as other household and respondent features.Since we expect respondents to be less accurate in averaging out off-take over 12 months the greater the day-to-day variability in off-take levels, the first variable in the regression is the household-specific coefficient of variation of the monitored off-take, computed as the standard deviation of total milk produced for all cows divided by its mean. We also include variables that reflect differences in management or milking practices that may be systematically related to recall quality: whether the household is in a village or camp, whether cows are milked once or twice per day, the number of cows milked, and the duration of the last lactation period. A variable measuring the number of cows that receive feed supplements is included, as this indicator can be related to both milk off-take per cow, as well as managerial ability or availability of resources on the part of the herder.

We hypothesize that respondents that are not exclusively focused on cattle rearing might recall events about livestock less accurately, and we use information on ownership of other animals, engagement in activities other than agriculture and source of income from migration as additional controls. The number of mobile phones owned by the household is included as a proxy for overall wealth, as well as ability to access and process information, while age of the household head is included on the grounds that ability to recall may decline with age. On the other hand, if younger farmers are less experienced, response accuracy could actually increase with age. Since two different enumerators collected data in the field, we also include a dummy to control for possible differences in enumerator's ability.

The most consistent, robust message that comes from this analysis is that the measurement error is correlated with the number of animals milked. The coefficient is of the expected positive sign, large in magnitude and highly statistically significant in five of the seven regressions we estimated. Interestingly, the coefficient is not significant only in the two methods based on the 6-month recall, suggesting that shortening the recall period may be an effective means to not only improve accuracy but also reduce bias.

Respondents living in the camps appear to be better able to recall the amount of milk off-take, and this is reflected in smaller measurement error. This may be linked to management practices, to the fact that livestock might be relatively more important in camps, and to reasons of ethnicity (Fulani herders are more likely to be residing in camps, compared to Zarma). It is hard to disentangle these effects and it should also be acknowledged that for reasons of collinearity, one should interpret with caution the regression coefficients that relate to management practices. This is the case for instance for the puzzling positive sign on the coefficient for the dummy capturing whether cattle are milked twice per day. We expected fewer milkings per day to be associated with better recall quality, but in fact it seems to be associated with greater measurement error.

The negative coefficient on the supplementation variable and positive coefficient on the duration of lactation variables, on the other hand, are expected, but only statistically significant for the lactation curve methods. We explain the former as reflecting greater managerial ability or simply greater importance given to animal management, and the latter to be related to the fact that the longer the milking period, the greater the degree of approximation implicit in the estimate of off-take employing the lactation curve method, and the related formula.

Turning our attention on some other possible household characteristics, we found little or no impact for the other household characteristics, which is not surprising given the relative homogeneity in the socio-economic composition of the villages studied. There are other two factors which we would have wanted to control for, namely the educational level and the gender of household head. Unfortunately, the level of education of the population in the district of Dantiandou is extremely low, even by Nigerien standards, and virtually our entire sample of households is headed by a man. In other settings these variables may however play a role. We take comfort in the result that measurement error does not appear to be influenced by the enumerator collecting the data.

TABLE 4 HERE

Finally, it is interesting to note how the overall fit of the lactation curve models is much higher compared to the other recall methods, whereas the simple 6-month recall has the lowest (with an adjusted R-squared equal to zero). That suggests that the lactation curve methods likely embed a larger degree of systematic error which correlates with several variables of interest related to livestock management, which is hardly a desirable feature when employing a productivity measurement in analytical work.

5. Conclusions

While there has been a renewed interested in the research over the nexus between agriculture, poverty and nutrition in recent years, associated with the increase in international food prices, this has not been matched by an improvement in the state of agricultural statistics. In Africa the availability and quality of agricultural sector data leave much to be desired, and that is particularly so for the livestock sub-sector. In terms of methods, livestock statistics offer peculiar challenges that are exemplified by the difficulties of collecting accurate milk off-take data at the household level. However, of the limited investments in livestock statistics, hardly any goes into methodological validation. The work documented in this paper takes its motivation from this state of affairs, and from the belief that given the abysmally low level of attention to this type of work, efforts to improve data quality can have substantial marginal returns and multiplier effects on research and policy analysis.

There are some clear messages we take away from work implemented in Niger to test different recall methods to capture household level milk off-take data, against a gold-standard of physical monitoring over a 12-month period.

The first is that even though there is a substantial amount of measurement error in the way even the best recall methods we tested perform in capturing household milk off-take, some methods do in fact perform fairly accurately, and much more accurately than what we expected when we designed this exercise. In particular, the methods do a reasonable job at estimating the more common central tendency measures (mean and median), as well as the distribution of milk off-take across sample households.

The methods that rank consistently better among those we compared are the 6-month AMD recall, and a 12-month AMD recall coupled with a lactation curve recall aid. The lactation curve method, on the other hand was consistently the worst performer, with differing patterns depending on the number of data points used to estimate the off-take level at different points in the lactation. Within the AMD method, the shorter recall period appears to significantly improve the estimates, even though it is uncertain the extent to which this result would hold if the 6-month recall interview were to be moved to another point in time, given the seasonality of milk off-take.

While we did find some evidence of the AMD method being more likely to return some extreme values (which is one of the perceived shortcomings of this method), this occurrence was rather limited in our sample and not frequent or large enough to undermine the overall performance of the method. In particular, not only did the AMD methods yield more accurate estimates of average milk off-take in our sample, but they also provided a more accurate depiction of the 'true' distribution, something that is as important when assessing the role of milk (and livestock) in general in livelihoods and attempting to capture the heterogeneity across households.

Another reason militating against the use of the lactation curve method and in favor of the AMD, is that the former seems to not only lead to larger measurement error, but also to a greater likelihood of measurement error being correlated to other variables of interest, such as herd size and length of the milking period, and hence of total milk off-take itself.

Last but not least, the LC method is in some ways more demanding on the respondent (who is prompted a few more questions) as well as the analyst, who needs to derive milk off-take estimates from the calculation of the area under the milk off-take curve as described in equation (1). To achieve the same result, the AMD method requires fewer questions, and a much simpler multiplication of daily average off-take times the length of the off-take period. On the other hand, it should be noted that we employed a variant of the LC method that implicitly asks the respondent to 'average out' off-take values across all the animals milked, whereas this method is often employed with reference to specific animals within a herd. While there are reasons to discard that approach in large scale national surveys with complex questionnaires, it should also be noted that this is a limitation of the results presented here, which do not provide a comparative evaluation of the LC method when implemented with reference to specific animals.

Another limitation of the study concerns its external validity, that is, the extent to which the conclusions that can be made based on our data apply to survey data collection in other areas in Sub-Saharan Africa or in other developing regions, and to animals other than dairy cattle. Both concerns can only be addressed by replicating similar methodological validation exercises in different settings. Ancillary evidence to the results presented in the paper do point to the fact that the distributions of the milk off-take estimates may perform very differently for large and small ruminants, due to the shorter lactation periods of the latter. But again, this speculation can only be verified through further research.

Taken together, the results presented in this paper have clear implications for future questionnaire design that we feel are strong enough to recommend using the better performing methods in future household surveys of small livestock keepers in extensive livestock systems in low-income settings, at least when the menu of workable options is one that includes the alternative we have tested. While there are limits to the external validity of these results, which should be repeated in different settings and for different species, we do maintain that the findings reported here are strong enough to be already taken up in future questionnaire design by National Statistical Offices, researchers, and anyone involved in household survey design.

Acknowledgements

The authors are grateful to all the individuals who took part to the monitoring and surveys in the village and camps of the Dantiandou district, together with the local authorities, for their patience and active participation to the year round monitoring and the repeated surveys. They are grateful to Moussa Kaka, to have toured and vaccinated the dairy cows, and to Oumar Moumouni and Seybou Garba for most ably and carefully conducting the interviews and monitoring of milk production. They would also like to thank Dr Mohamadou Gandah and his colleagues from the ICRISAT administration office for the efficient logistical support they provided to the field team. Alberto Zezza is indebted to Derek Baker, Isabelle Baltenweck, Gero Carletto, Celine Dutilly, Mathieu Lesnoff, Nancy McCarthy, Jane Poole, Steve Staal, and Matthew Turner and for guidance, discussion, and comments at various stages of this project. Comments from two anonymous reviewers were very helpful in improving an earlier draft of the paper. This work was supported by the Livestock Innovation in Africa Project, funded by the Bill and Melinda Gates Foundation.

REFERENCES

Beegle, K., De Weerdt, J., Friedman, J., Gibson, 2012. “Methods of household consumption measurement through surveys: Experimental results from Tanzania”. *Journal of Development Economics* 98 (1): 3-18. dx.doi.org/10.1016/j.jdeveco.2011.11.001

Beegle, K., Carletto, G., Himelein, K., 2011. “Reliability of Recall in Agricultural Data”. World Bank Policy Research Working Paper 5671. 5671. dx.doi.org/10.1016/j.jdeveco.2011.09.005

Bocoum, Ibrahima; Yahaya, Atte Issa et Sidi; Zezza, Alberto. 2013. “L'elevage et les conditions de vie des menages au Niger : une analyse descriptive de l'enquete sur les conditions de vie des menages et l'agriculture”. Washington DC ; World Bank Group.

Carletto, G., 2009. “Improving the Availability, Quality and Policy Relevance of Agricultural Data: The Living Standards Measurement Study – Integrated Surveys on Agriculture”. Development Research Group, the World Bank, Washington, DC.

Carletto, G., Savastano, S., Zezza, A., 2013. “Fact or Artefact: The Impact of Measurement Errors on the Farm Size-Productivity Relationship”. *Journal of Development Economics,* 103: 254–261. dx.doi.org/10.1016/j.jdeveco.2013.03.004

Davis, B., Winters, P., Carletto, G., Covarrubias, K., Quinones, E. J., Zezza, A., Stamoulis, K., Azzarri, C., Di Giuseppe, S., 2010. “A Cross-Country Comparison of Rural Income Generating Activities”. *World Development* 38 (1): 48-63.

De Mel, S., McKenzie, D.J., Woodruff, C., 2009, “Measuring microenterprise profits: Must we ask how the sausage is made?”, *Journal of Development Economics*, 88 (1): 19–31.

Deininger, K., Carletto, G., Savastano, S., Muwonge, J., 2012. “Using diaries to improve crop production statistics: Evidence from Uganda”. *Journal of Development Economics,* 98 (1): 42-50. dx.doi.org/10.1016/j.jdeveco.2011.05.007

Delgado C., M. Rosegrant, H. Steinfeld, S. Ehui, and C. Courbois. 1999. “Livestock to 2020: The Next Food Revolution”. *Food Agriculture, and Environment Discussion Paper* 28. International Food Policy Research Institute.

Dewey, K., 2005. “Guiding Principles for Feeding Non-Breastfed Children 6-24 months of Age”. Geneva, World Health Organization.

Dillon, A., Bardasi, E., Beegle, K., Serneels, P., 2012. “Explaining variation in child labor statistics”. *Journal of Development Economics,* 98 (1): 136-147. dx.doi.org/10.1016/j.jdeveco.2011.06.002

Dror, D.K., Allen, L.H., 2011. “The importance of milk and other animal-source foods for children in low-income countries”. *Food and Nutrition Bulletin,* 32 (3): 227-243.

FAOSTAT, 2015. FAOSTAT Online database, accessed 3 June 2015.

Gerosa, S., Skoet, J. (2013). “Milk availability: Current production and demand and medium-term outlook”. In Muehlhoff, E., Bennett, A., McMahon, D. (eds.) *Milk and dairy products in human nutrition*. FAO Publications, Rome.

Gibson, J., Beegle, K., De Weerdt, J., Friedman, J., 2013. “What Does Variation in Survey Design Reveal about the Nature of Measurement Errors in Household Consumption?”. World Bank Policy Research Working Paper 6372. 6372. DOI: 10.1111/obes.12066.

Hoddinott, J., Headey, D., Dereje, M., 2014. “Cows, missing milk markets and nutrition in rural Ethiopia”. Ethiopia Strategy Support Program Working Paper 63, IFPRI: Washington, DC.

Hoppe, C., Mølgaard, C. and Michaelsen, K.F., 2006. “Cow's Milk and Linear Growth in Industrialized and Developing Countries”. *Annual Review of Nutrition,* 26: 131-173.

Iarossi, G., 2006. “The Power of Survey Design. A User’s Guide for Managing Surveys, Interpreting Results, and Influencing Respondents”. World Bank: Washington, DC.

Joliffe, D., 2001. “Measuring absolute and relative poverty: The sensitivity of estimated household consumption to survey design”. *Journal of Economic and Social Measurement,* 27 (1): 1-23.

Lesnoff, M., Messad. S, Juanes, X., 2010. 12MO. “A cross-sectional retrospective method for estimating livestock demographic parameters in tropical small-holder farming systems”. CIRAD: Montpellier, France.

Muehlhoff, E., Bennett, A., and McMahon, D. (eds.), 2013. *Milk and dairy products in human nutrition*. FAO Publications, Rome.

Njuki, J., Poole, J., Johnson, N., Baltenweck, I., Pali, P., Lokman, Z., Mburu, S., 2011. “Gender, livestock and livelihood indicators”. ILRI, Nairobi.

Pica-Ciamarra, U., Tasciotti, L., Otte, J., Zezza, A., 2011. “Livestock assets, livestock income and rural households. Cross-country evidence form household surveys”. FAO, Rome.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M., Singer, E, 2004a. “Methods for Testing and Evaluating Survey Questions”. *Public Opinion Quarterly*, 68(1): 109-130.

Presser, S., M.P. Couper, J.T. Lessler, E. Martin, J. Martin, J.M. Rothgeb, E. Singer (eds.). 2004b. *Methods for Testing and Evaluating Survey Questions*. Wiley.

Roberts, John M., Brewer, Devon D., 2001. “Measures and tests of heaping in discrete quantitative distributions”. *Journal of Applied Statistics*, 28 (7): 887–896. DOI:

10.1080/02664760120074960.

Sadler, K., Kerven, C., Calo, M., Manske, M., Catley, A., 2009. “Milk Matters: A Literature Review of Pastoralist Nutrition and Programming Responses”. Feinstein International Center, Tufts University and Save the Children, Addis Ababa.

Scott, C., Amenuvegbe, B., 1990. “Effect of Recall Duration on Reporting of Household Expenditures”. World Bank Social Dimensions of Adjustment in Sub-Saharan Africa Working Paper No. 6.

Thornton, P. K., Kruska, R. L., Henninger, N., Kristjanson, P. M., Reid, R. S., Atieno, F., Odero, A., Ndegwa, T., 2002. “Mapping poverty and livestock in the developing world”. International Livestock Research Institute, Nairobi, Kenya.

Weaver, C., Wijesinha-Bettoni, R., McMahon, D., Spence, L., 2013. “Milk and dairy products as part of the diet”. In Muehlhoff, E., Bennett, A., and McMahon, D. (eds.). *Milk and dairy products in human nutrition*. FAO Publications, Rome.

Wijesinha-Bettoni, R., Burlingame, B., 2013. “Milk and dairy product composition”. In Muehlhoff, E., Bennett, A., and McMahon, D. (eds.). *Milk and dairy products in human nutrition*. FAO Publications, Rome.

Wilson, T., Pearson A., Bradbear, N., Jayasuriya, A., Laswai, H., Mtenga, L., Richards, S., Smith, R., 2005. “Livestock products – Valuable and more valuable”. In: Owen, E. A., Kitalyi, A., Jayasuriya, N., Smith, T. (eds.). *Livestock and Wealth Creation: Improving the Husbandry of Animals Kept by Resource-Poor People in Developing Countries*. Nottingham University Press, UK.

Tables

Table 1: *Summary statistics for different randomly selected sub-samples*

		Avg. raised cows	Avg. lactating cows	Months of milking	Avg. cows milked	Length of previous lactation	Gap in last two births	Age of cow at first birth	Number of births	Age of cow	
questionnaire type // unit of measurement		*units*	*units*	*months*	*units*	*months*	*months*	*months*	*units*	*years*	
Lactation Curve Quest.	*Obs.*	172	170	168	168	175	155	168	169	168	
	Mean	5.64	2.73	11.07*	2.02	12.35	22.17	52.23	3.04	10.55	
	Median	4	2	12	2	12	24	60	3	10	
	Std. Dev.	5.09	1.84	2.31	1.19	4.36	5.31	24.46	1.39	3.14	
	Min	1	1	2	1	4	12	4	1	5	
	Max	31	10	12	7	30	36	108	7	19	
Avg. / Herd / Day (AMHD) Quest.	*Obs.*	168	164	163	163	157	154	164	165	164	
	Mean	5.88	2.76	10.50*	2.02	12.61	22.30	52.23	2.98	10.27	
	Median	4	2	12	2	12	24	60	3	10	
	Std. Dev.	5.51	2.08	2.98	1.18	4.85	5.43	23.98	1.45	3.04	
	Min	1	1	1	1	3	12	3	1	5	
	Max	35	12	12	7	28	36	96	9	18	
Lactation Curve Quest.	*Obs.*	135	135	135	135	138	122	133	134	133	***Monitored***
	Mean	5.93	2.82	11.13	2.01	12.46	22.11	51.62	3.05	10.45	
	Median	5	2	12	2	12	24	60	3	10	
	Std. Dev.	5.16	1.90	2.21	1.11	4.36	5.26	24.45	1.37	3.03	
	Min	1	1	2	1	5	12	4	1	5	
	Max	31	10	12	6	30	36	96	7	19	
Avg. / Herd / Day (AMHD) Quest.	*Obs.*	134	134	134	134	127	125	134	134	134	
	Mean	6.25	2.91	10.69	2.05	12.34	21.96	51.77	3.02	10.38	
	Median	4.5	2	12	2	12	23	60	3	10	
	Std. Dev.	5.81	2.18	2.82	1.17	4.63	5.40	24.33	1.48	3.05	
	Min	1	1	1	1	3	12	4	1	6	
	Max	35	12	12	7	28	36	96	9	18	
Lactation Curve Quest.	*Obs.*	37	35	33	33	37	33	35	35	35	***Not monitored***
	Mean	4.57	2.37	10.82	2.03	11.95	22.42	54.57	3.00	10.94	
	Median	3	2	12	2	12	22	60	2	10	
	Std. Dev.	4.73	1.57	2.69	1.47	4.39	5.54	24.71	1.48	3.56	
	Min	1	1	2	1	4	12	5	1	5	
	Max	25	7	12	7	24	36	108	7	17	
Avg. / Herd / Day (AMHD) Quest.	*Obs.*	34	30	29	29	30	29	30	31	30	
	Mean	4.41	2.10	9.62	1.90	13.77	23.76	54.27	2.81	9.80	
	Median	3	2	12	2	12	24	60	2	10	
	Std. Dev.	3.85	1.37	3.58	1.23	5.66	5.42	22.64	1.30	3.00	
	Min	1	1	1	1	4	12	3	1	5	
	Max	14	5	12	5	24	36	96	7	17	

*Source: Authors' calculation based on data collected for the experiment. Significance levels: * 10%; ** 5%; *** 1%.*

Table 2: *Household milk off-take (liters). Comparison of monitoring and recall data, various methods (annual and 6-month).*

	TOTAL						VILLAGE						CAMP					
	Obs.	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**	**Obs.**	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**	**Obs.**	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**
Physical monitoring																		
Monitoring at 12 months	300	877	741	631	10	3291	129	605	512	465	10	2484	171	1083	971	662	45	3291
Recall on 12 months																		
6 months recall - annualized	171	847	720	699	8	3600	63	569	360	534	8	3240	78	1089	1080	640	180	2880
Avg. / Herd / Day (AMHD) - LC module	167	934	720	870	43	5400	55	684	557	591	43	2229	79	1072	929	845	130	4458
Avg. / Herd / Day (AMHD) - All	330	926	720	863	9	6687	111	759	557	692	9	3960	157	1049	743	880	130	6687
Avg. / Herd / Day (AMHD) - 12 months recall	163	918	720	859	9	6687	56	832	549	777	9	3960	78	1027	743	920	130	6687
Lactation curve - 3 points	167	1200	818	1146	132	6900	56	1091	600	1140	132	6037	79	1284	913	1229	201	6900
Lactation curve - 4 points	167	990	693	934	87	6037	56	915	480	1010	87	6037	79	1055	855	954	174	5263
	Note: 330 sample of households with not null milk production.																	
Monitoring at 6 months	300	471	386	323	10	1825	129	334	267	230	10	1321	171	574	509	345	45	1825
Recall at 6 months	171	424	360	350	4	1800	63	284	180	267	4	1620	78	545	540	320	90	1440
	Note: 300 monitored hhs (152 with LC quest. / 148 with AMHD quest.).																	
Source: Authors' calculation based on Dantlait survey data.																		

Table 3: Correlation and regression (Ordinary Least Squares, OLS) coefficients between monitoring and recall methods

	Correlation coefficient	OLS no constant		OLS		OLS (logs)		N
Correlation with 12 months monitoring		**Coeff**	**R2**	**Coeff**	**R2**	**Coeff**	**R2**	
6 months recall - annualized	0.71	0.91	0.81	0.68	0.50	0.76	0.63	141
Avg. / Herd / Day (AMHD) - LC module	0.61	0.79	0.72	0.51	0.38	0.57	0.48	134
Avg. / Herd / Day (AMHD) - All	0.52	0.73	0.66	0.41	0.27	0.58	0.44	268
Avg. / Herd / Day (AMHD) - 12 months recall	0.44	0.69	0.60	0.33	0.19	0.58	0.41	134
Lactation curve - 3 points	0.35	0.47	0.52	0.19	0.12	0.47	0.21	135
Lactation curve - 4 points	0.36	0.57	0.53	0.24	0.13	0.49	0.24	135
Correlation with 6 months monitoring								
Recall at 6 months	0.67	0.97	0.78	0.69	0.44	0.76	0.63	141

Source: Authors' calculation based on Dantlait survey data.

Table 4: Regressions' results on the determinants of the measurement errors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	6-month recall annualized	Avg / Herd / Day - LC module	Avg / Herd / Day - All	Avg / Herd / Day - 12-month recall	Lactation Curve 3-point	Lactation Curve 4-point	Recall at 6 months
Coefficient of variation	-0.514	-0.127	0.468	0.463	-0.986	-0.527	-0.253
	(0.317)	(0.551)	(1.075)	(1.698)	(1.148)	(0.960)	(0.230)
Dummy Territory (1=Camp 0=Village)	0.121	-0.404*	-0.393	-0.233	-0.780**	-0.618**	0.252
	(0.229)	(0.209)	(0.323)	(0.486)	(0.339)	(0.279)	(0.182)
Dummy =1 if milk is collected only in the morning	0.290	0.367*	0.680**	0.974*	0.430	0.405	0.250*
	(0.186)	(0.195)	(0.303)	(0.574)	(0.323)	(0.290)	(0.137)
Enumerator dummy	-0.119	-0.016	-0.036	-0.108	0.073	-0.107	-0.073
	(0.186)	(0.173)	(0.234)	(0.473)	(0.376)	(0.318)	(0.135)
Log number of cows	-0.106	0.677***	0.783***	0.962***	1.232***	1.143***	-0.033
	(0.142)	(0.190)	(0.204)	(0.327)	(0.380)	(0.371)	(0.120)
Number of supplemented cows	-0.052	-0.040	-0.046	-0.068	-0.258***	-0.222***	-0.003
	(0.039)	(0.050)	(0.033)	(0.053)	(0.079)	(0.069)	(0.032)
Log of age of hh head	-0.095	0.510**	0.235	0.126	0.405	0.423	0.141
	(0.277)	(0.242)	(0.355)	(0.664)	(0.514)	(0.449)	(0.180)
Annual household remittances received (1,000 CFA)	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Dummy =1 if agriculture is the only hh activity	0.028	-0.391*	-0.370*	-0.410	0.013	0.048	0.036
	(0.151)	(0.229)	(0.201)	(0.347)	(0.339)	(0.310)	(0.126)
Number of mobile phones owned	0.102	0.040	-0.027	-0.192	0.151	0.182	0.060
	(0.063)	(0.098)	(0.077)	(0.157)	(0.195)	(0.198)	(0.056)

Number of other animals	0.051	0.021	-0.050	-0.165	0.018	-0.003	0.023
	(0.047)	(0.051)	(0.053)	(0.112)	(0.069)	(0.062)	(0.039)
Duration of previous lactation	0.002	-0.017	-0.041	-0.051	0.175***	0.142**	-0.006
	(0.015)	(0.024)	(0.028)	(0.041)	(0.060)	(0.057)	(0.010)
Constant	0.511	-1.523	-0.232	0.726	-2.418	-2.711	-0.743
	(1.068)	(1.187)	(1.384)	(2.225)	(2.485)	(2.386)	(0.689)
Observations	129	134	266	132	135	135	129
Log – likelihood	-146.6	-181.9	-533.9	-299.1	-254.8	-240.2	-115.2
Prob. > F	0.773	0.000	0.001	0.124	0.000	0.000	0.413
R-squared	0.105	0.199	0.091	0.100	0.370	0.340	0.094
Adj. R-squared	0.013	0.120	0.048	0.009	0.308	0.275	-0.000
RMSE	0.795	0.990	1.846	2.457	1.681	1.508	0.623

Note: Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figures

Figure 1: *Computing milk off-take using the LC method*

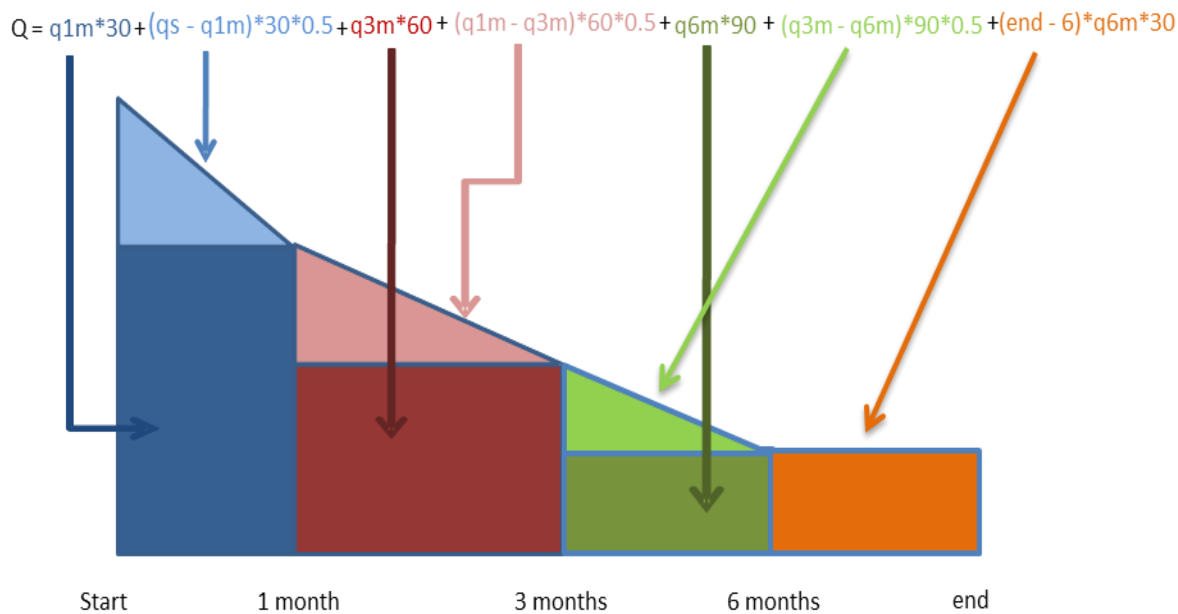


Figure 2: *Comparison of recall methods*

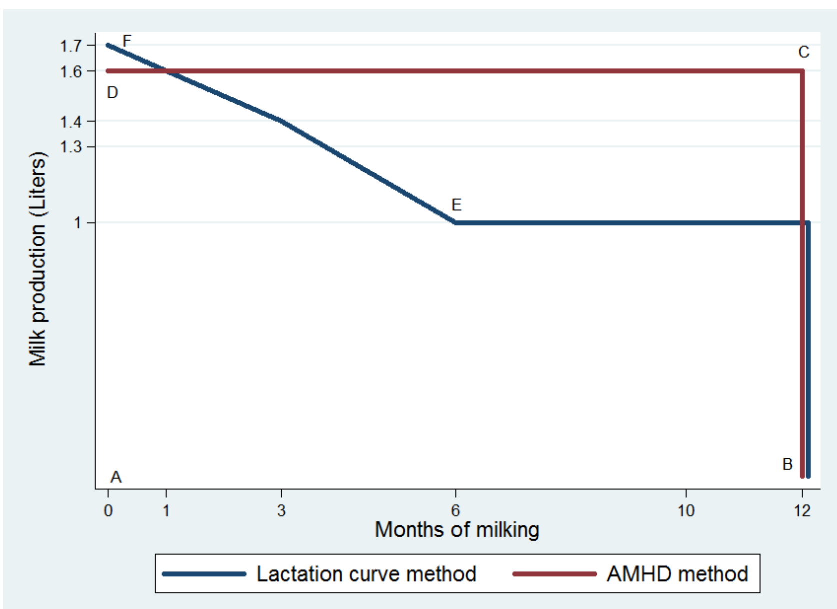
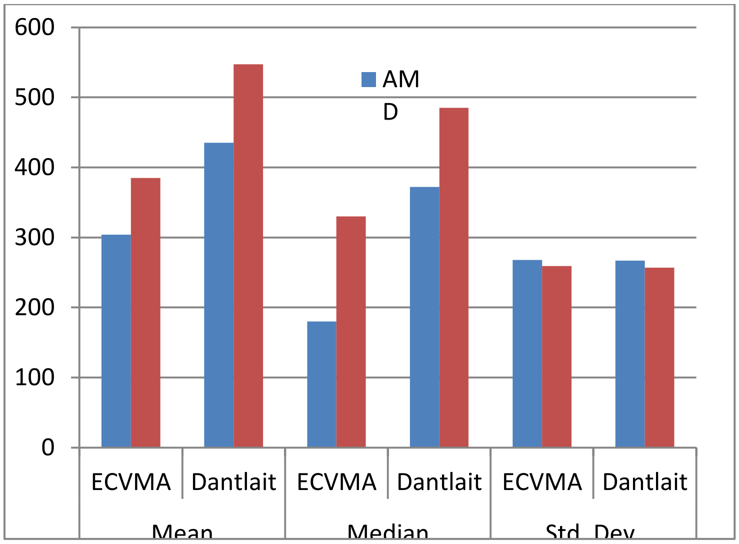
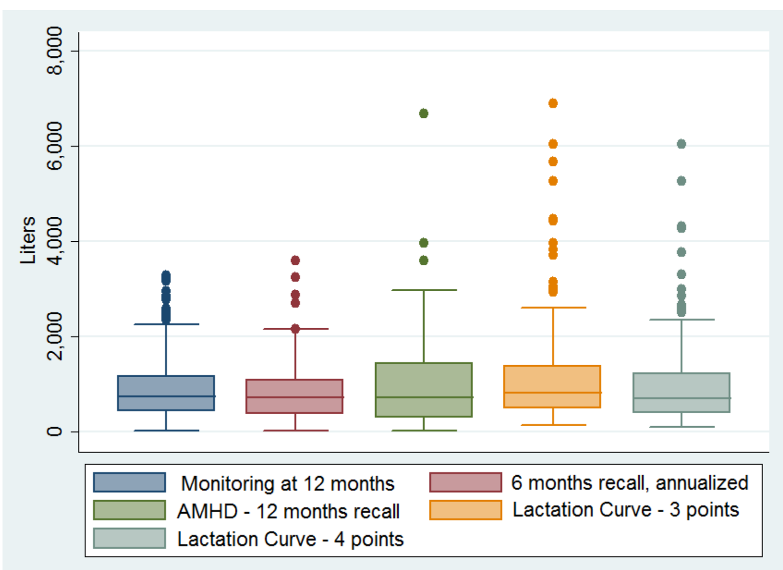


Figure 3. *Comparison of mean, median and standard deviation measures of milk off-take estimates from AMD and LC methods in Dantlait and ECVMA surveys (liters)*



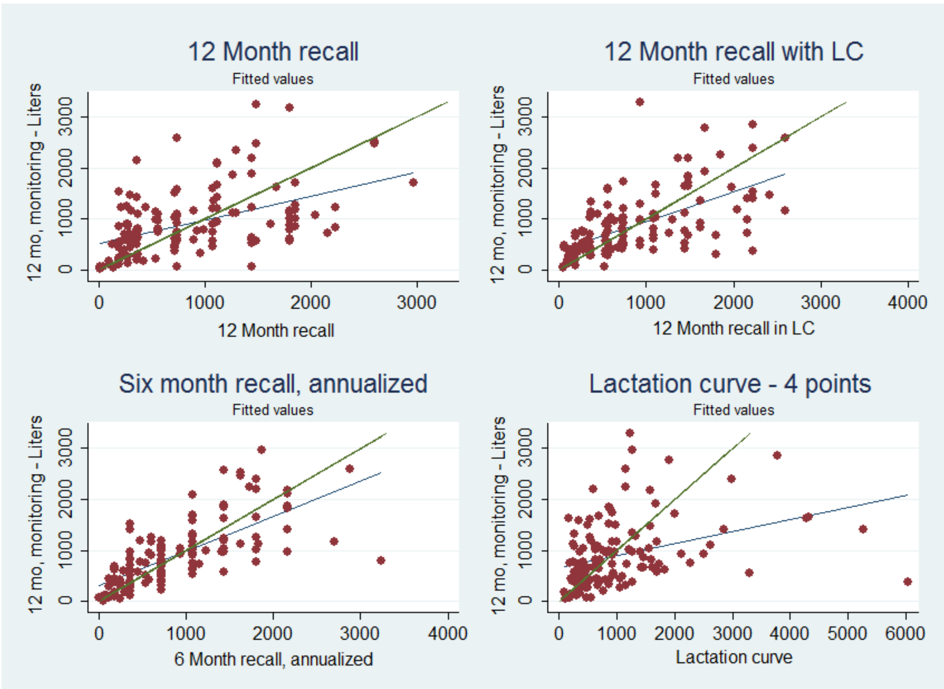
Source: Dantlait and ECVMA surveys

Figure 4: *Box plots of mean household daily milk off-take (liters): Monitoring and recall*



Source: Dantlait survey

Figure 5: *Household milk off-take (liters): Scatter plots of recall against monitoring method*



Source: Dantlait survey