How much should we trust micro-data? A comparison of the socio-demographic profile of Malawian households using LSMS and DHS data

Luca Tasciotti[a,b,c] and Natascha Wagner[a]

[a]International Institute of Social Studies of Erasmus University Rotterdam, Kortenaerkade 12, 2518 AX Den Haag, Netherlands
[b]School of Oriental and African Studies (SOAS) Thornhaugh Street, London WC1H 0XG, United Kingdom

[b]Corresponding author: tasciotti@iss.nl

Abstract

This paper assesses the empirical representativeness of micro-data by comparing the Malawi 2008 census to two representative household surveys – the Living Standard Measurement Survey' and the 'Demographic and Health Survey' – both implemented in Malawi in 2010. The comparison of descriptive statistics –demographics, asset ownership and living conditions– shows considerable similarities despite statistically identifiable differences due to the large samples. Differences mainly occur when wording, scope and pre-defined answer categories diverge across surveys. Multivariate analyses are considerably less representative due to loss of observations with composite indicators yielding higher comparability as individual ones. Household level fixed-effects specifications produce more similar results yet are not suited for policy conclusions. Comparability of micro-data should not be assumed but checked on a case-by-case basis. Still, micro-data constitute reliable grounds for factually informed conclusions if design and context are appropriately considered.

Keywords: Household data, survey comparison, representativeness, Sub-Saharan Africa, Malawi
JEL: C80, C83, D00, O12

1. Introduction

The collection of large amounts of data –commonly used to describe economic performance, test theories and assess the impact of projects– represents a major achievement. Nowadays, the great majority of the policy recommendations made by economists are based on prior quantitative analyses (Heckman, 2001).

Especially in developing countries, where administrative and routinely collected data are scarce, micro-surveys constitute important tools to collect information about the state of an economy. The exponential proliferation of data collections in developing countries has greatly contributed to the surge of evidence-based policymaking (Ravallion, 2003). The majority of current research in development economics relies on the use of survey data either collected by institutions using multi-purpose survey instruments or collected in the context of programme evaluations (Ravallion and Chen, 1997). The general assumption underlying the use of these data is that they represent the reference population with a high degree of accuracy (Holt, 1985; Elbers et al, 2003).

However, in light of their high demand large-scale data collections have been criticized for possibly being incomplete due to missing respondents or responses, imperfect due to differences in perceptions, inaccurate due to misconceptions at either side of the interview process, and potentially non-representative due to ad-hoc convenience sampling. As a response, data collectors use standard statistical procedures and broadly-accepted, widely-used questionnaires, and have developed hands-on manuals to instruct enumerators. Consequently, there is no longer doubt about the rigor and accuracy of the *statistical* sampling procedures. In this paper we assess the *empirical* similarity of two different micro-data collections from the same country and year by comparing them with each other and population census data.

Studies similar to the current one have been carried out with the objective of detecting data quality problems and discrepancies between data from different sources. Srinivasan (1994) points out that the statistical office in India used to produce three different indicators for the production of food crops, which differed in their levels and in their trends. Grosh and Glewwe (1996) also raised the issue of data quality; among others they identified country coverage and quality as challenges for the comparability of household surveys. Sandefur and Glassman (2015) look at educational and health related indicators managed by local administrative agencies versus those computed by international agencies via independent household surveys, i.e. the Demographic and Health Survey. The study, carried out across multiple African countries, concludes that official statistics systematically overestimate development progress. The inconsistency problem between household surveys and national accounts has further been highlighted by Deaton (2005); in both India and the United States, per capita consumption based on household survey data rises roughly 1% per year more slowly compared to estimations using national accounts data. Deaton concludes by advocating an ‘international initiative to provide a set of consistent international protocols for survey design, as well as deeper study of the effects of non-sampling errors, particularly non-compliance.’

This study focuses on Malawi as it was ranked the poorest country in the world in 2015 based on 2010–2014 GDP per capita (World Bank, 2015). We analyze data from the 2008 population census, the Living Standard Measurement Survey (LSMS) and the Demographic and Health Survey (DHS), both carried out in 2010. The three data collections focus on slightly different aspects, with the census putting an emphasis on demographic dynamics and housing, the LSMS zooming in on income generating activities, and the DHS concentrating on health-related matters. The census represents an official enumeration of the population. For the LSMS and the DHS, sample size calculations and sampling procedures were employed. Sampling was based on the enumeration areas established in the census and comparable across the two surveys. However, from the *empirical* point of view, the LSMS and DHS do not necessarily represent the same population or the same aspects of the population’s livelihoods when comparing the *actual data* that was gathered. For those indicators, collected both in the census and the two surveys, we demonstrate that the demographic composition and descriptive statistics of key assets and indicators of living conditions show considerable similarity. Yet, even minor differences in average statistics show up statistically significant because our analysis relies on large samples. Nevertheless, we do not consider the majority of the differences as economically meaningful because they are not large enough to suggest that an important divergence was detected across surveys. In the demographic statistics, we find less than 3 percentage points deviations between the 2008 census and the two surveys except for one comparison where the deviation is 4.27 percentage points. The use of firewood and toilet facilities is documented with hardly any discrepancies across surveys. There are some moderate differences for drinking water, which are most likely linked to variations in the pre-defined answer categories across surveys. More pronounced differences mainly occur when the wording and the scope of the questions differ. This highlights that the survey instruments are not neutral; the way the questions are formulated and pre-determined answer categories are defined influence the findings and resulting conclusions. Surprisingly, we also find differences in the ownership of key assets suggesting that even the possession of countable goods can be documented according to different standards. Nevertheless, the descriptive statistics from the survey data are overall not only highly comparable but also representative as shown by the comparison with the 2008 census.

Multivariate analyses are markedly less representative due to loss of observations resulting in a subsample analysis rather than the study of a fully representative sample. For the LSMS and the DHS we establish correlations between two human capital indicators (child health and education) and the household demographic and asset indicators identified in the descriptive analysis. There was no similarly detailed information available in the 2008 census. The findings suggest that multivariate correlations among human capital and asset indicators are very susceptible to the objective of the original survey questions. The LSMS-based results suggest that children from livestock keeping households have higher levels of schooling (Randolph et al, 2007), whereas the DHS-based results indicate that mobile phone technology and the quality of the housing influences this more prominently (Yuyu and Li, 2009). Concerning educational outcomes, the LSMS data suggest that children and grandchildren of household heads perform considerably better with coefficient estimates being three times larger as compared to the DHS results. The DHS-based results, in turn, point out the correlation between a proper water and sanitation infrastructure and a higher level of schooling (Checkley et al, 2004). Thus, potential policy conclusions based on multivariate regressions may differ substantially according to the dataset employed. This is not as much the case if our analysis focuses on a demanding household fixed-effects specification. Here, we identify more similar individual-level correlates across survey datasets. While rigorous and displaying more coherent results, however, such models cannot be employed to answer policy questions about household or community dynamics.

Lastly, we also show that differences in policy conclusions are likely not to be driven by survey aspects that are comparable but rather by inconsistencies within and across surveys and indicators. DHS only allows for the construction of within survey wealth-quintiles whereas LSMS is used to construct poverty profiles based on national poverty lines. Moreover, comparability between DHS and LSMS is limited to general categories, say of ownership, often expressed as dummy variables. But once we are interested in the quality and characteristics of the assets owned analyses have to rely exclusively on one dataset. Thus, the comparability of micro-data should not be assumed but should be checked on a case-by-case basis.

The remainder of the paper is structured as follows. Section 2 discusses the potential sources of bias in micro-datasets. Section 3 presents background information about the Malawian 2008 census and the 2010 LSMS and DHS. Section 4 compares descriptive statistics from the census with the two surveys. Multivariate correlations analyzing the determinants of child health and schooling are presented in Section 5. Although not the main focus of the paper, in section 6 we briefly discuss incomparability and divergency issues across surveys. Section 7 concludes with the implications of our findings and resulting recommendations.

2. Potential sources of bias in survey data

International organizations (e.g. World Bank, International Monetary Fund or Food and Agriculture Organization) and large non-governmental organizations use statistical data collected in less developed countries to underpin their work. The quality and reliability of the resulting micro-data is an important concern as those data often constitute the basis for policy briefs and government recommendations, in particular for the poorest countries such as Malawi.

Micro- and macro-data collection emerged prominently in the last few decades. In 1962, Orcutt proposed a micro-macro data combination to get a more credible description of underlying economic phenomena and to test alternative economic theories. One year later, Morgenstern (1963) made a comprehensive review arguing that the analysis of unreliable and biased data lead to distorted policy conclusions with dramatic consequences for the welfare of developing countries. Heckman (2001) further illustrated the problem of data accuracy adding examples from India, which has one of the best statistical systems in the developing part of the world.

The quality of survey data can be evaluated along five dimensions of bias. First, the surveyed samples' representativeness of the population. A standard procedure is to apply random sampling to reduce sampling error, which is caused if only a subset of the population is considered for the survey instead of the whole population. The resulting error is the difference between the sample statistic and the actual, but unknown, population parameter. The procedure can be introduced, for example, when absent households are replaced for convenience without further trying to track them. A related concept is 'selection bias', which arises when only a non-random subset of the reference population can be observed or is purposely sampled. The resulting statistics and estimated structural models only represent the particular sample for which the results are derived and no general conclusions about the entire population can be drawn. For random samples, larger groups of respondents will increase the accuracy of the collected data, with the optimal sample size being calculated based on the variability of the issue of interest (Deaton, 1997).

Second, the comparability of survey data over time and across countries is not necessarily given. Survey questions, methods of measurement, recall periods, and indicator definitions are not the same across countries leading to inconsistent results in cross-country or panel-data studies (Srinivasan, 1994). For example, comparisons of the 'dollar a day' poverty rates show that some surveys use income to measure the well-being of a given household while others rely on consumption measures rendering comparisons questionable (Ravallion, 2003).

Third, most often it is assumed that measurement errors are *not systematic* but purely random (Bound et al, 2001). Measurement errors reduce the efficiency of multivariate analyses whenever the dependent variable is measured with error. Bias is introduced when any of the independent variables suffers from measurement error. Especially income, expenditures, and subjective indicators are susceptible to measurement errors due to recall bias and differences in perceptions (Alwin, 1989), whereas variables such as level of schooling and asset ownership can be collected with reasonable accuracy. Systematic measurement errors can be introduced when one or more survey enumerators misinterpret a question due to poor wording or low levels of literacy. The impact of measurement errors on sample statistics and estimated parameters depends on the magnitude of the errors and its correlation with the true variable (Bound et al, 2001).

Fourth, even if survey respondents are properly identified and interviewed, they might refuse or be unable to answer questions about their income or past events, which results in missing data due to non-response (Brick and Kalton, 1996). This reduces the original sample size, and threatens representativeness and statistical power (Schoumaker, 2011). Ultimately, missing data also lead to bias in the survey estimates if they are simply ignored (Langkamp et al, 2010). Distortions can be considerable in multivariate analyses when all those observations displaying missing values across variables are dropped from the analysis. Imputation, while not always recommended, is often used to replace missing survey responses (Little and Rubin, 2014).

Fifth, non-sampling errors are mainly associated with data collection and processing procedures. They include the characteristics of the interviewer (e.g. age, gender, race), the intrinsic characteristics of the reference population, the topic investigated by the survey, the design and administration of the questionnaire (e.g. face-to-face, telephone, self-administered), and the specific conditions of measurement.

In conclusion, reliable, high quality data can only be collected if surveys are carefully designed and executed. Enumerators and analysts need to be aware of the potential sources of bias to be able to minimize them. Comparability, reliability, and representativeness of micro-datasets theoretically hold if the sources of bias are minimized, if not eliminated (Griliches, 1986). Nowadays, more than a single source of data for the same country, time period or phenomenon is often available. Validation and cross-checking of data from different sources allow for a good description of underlying population dynamics (Sutherland et al, 2002). In the following sections, we proceed with such an empirical validation for the case of Malawi.

3. The set-up of the Malawian census, LSMS and DHS

This section briefly outlines the main data collection features of the 2008 census and the 2010 LSMS and DHS data collections. All three data collections were implemented by the Malawian National Statistical Office (NSO) in collaboration with different partners. The Population and housing census was conducted in June 2008. It was the fifth in a series of decennial censuses since the country's independence in 1964. Census activities were based on the demarcation of 12,631 enumeration areas, and carried out by more than 10,000 enumerators (National Statistical Office, 2008).

Both the LSMS and DHS rely on the enumeration areas as defined in the 2008 census, and adopted a two-stage sampling with probability proportional to size and stratification into rural and urban areas within districts. The LSMS and the DHS collected cross-sectional data but the LSMS constitutes the baseline for a panel subsample follow-up. The major, obvious difference between the two surveys is the sample size. The LSMS consists of 768 primary sampling units (PSU) with 16 households within each PSU resulting in a total sample of 12,271 households and attaining district level representativeness. The DHS includes almost 100 PSUs more with 20 urban and 35 rural households per PSU, resulting in a nationally representative sample of 27,307 households. Another difference is represented by the timing. The LSMS took place between March 2010 and March

2011 whereas the DHS was carried out between June and November 2010.

The 2010 LSMS was the third integrated household survey carried out in Malawi in collaboration with the World Bank. The first LSMS was implemented in 1997 and since then a survey has been carried out approximately every five years. The 2010 LSMS has a particular focus on income generating activities, time use, labor, household enterprises, agricultural/fishing activities, expenditures for food/non-food items and asset ownership. The survey also contains modules about education, health and child anthropometrics (National Statistical Office, 2012).

The 2010 DHS was the fourth DHS following surveys conducted in 1992, 2000 and 2004. Its main focus is on demographic characteristics and health. Therefore, the survey builds around elaborate health questions related to anthropometrics, vitamin A, tuberculosis, HIV, malaria and anemia testing. The survey only inquires ownership of some key assets but does not contain a detailed income and labor module.

4. Comparison of descriptive statistics from the 2008 census, LSMS and DHS data

For the comparison of descriptive statistics we rely on those indicators that are found in both survey dataset. Whenever similar census data is available, we also draw the comparison with the 2008 census to assess the representativeness of the survey data. We present the demographic composition, household head characteristics, the overall housing and living situation, and asset ownership. For the LSMS and DHS data we apply the population weights as provided by the surveys.

We assess both the statistical and economic differences across datasets. By economic difference we refer to the magnitude of the discrepancy that is assessed relative to the average level. This means that small differences, at low levels, can be economically more meaningful compared to moderate differences at higher levels.

In Table 1 we present the demographic composition of the three datasets. We notice a similar gender mix with differences across datasets amounting to less than 0.5 percentage point. The population age composition is most comparable across the two surveys with differences being smaller than 2 percentage points except for the youngest age cohort. Differences are slightly more pronounced when we compare the surveys with the census data. Yet, they still amount to less than 5 percentage points and are again most pronounced for the younger cohorts where transitions from one cohort to another are likely to occur within the 2-year window from 2008 to 2010. Thus, in terms of age composition the surveys capture similar populations and are representative according to the 2008 census.

Turning to household characteristics (Table 2), we find that the average age of the household head ranges between 42.3 (LSMS) and 42.7 (DHS) years. The small difference in age shows up statistically significant due to the large sample sizes of the datasets even if the actual difference of 0.45 year is rather moderate. The share of male household heads is, with more than 70%, equally high in all three datasets with the census indicating a share of 73%, which lies between the LSMS and DHS. While there is a statistical divergence between the LSMS and DHS in the share of household heads, the difference of 4.3 percentage points is small in magnitude relative to the high-base level of approximately 70%.

Less than 10% of the households have access to electricity with the DHS dataset suggesting a 1.5 percentage points higher coverage while LSMS and the census data are aligned. This difference across the two surveys, albeit small, is economically interesting because every additional household that receives access to electricity presents a non-negligible progress at these low levels of electrification. In this case, the DHS might slightly over-represent electrified households. Furthermore, the DHS suggests that the average household has 0.63 (0.68) fewer rooms compared to LSMS (2008 census). The difference between the DHS and LSMS stems from the way the survey question was posed. The DHS asked ‘How many rooms are used for sleeping?’ whereas the LSMS asked ‘How many separate rooms do the members of your household occupy (excluding bathrooms, toilets, storerooms, or garage)?’ The variable ‘number of rooms’ is a good example showing that analysts need to know the precise wording of the survey questions in order to draw appropriate policy conclusions. In the DHS, a separate question about kitchen facilities consists of more than 90% missing responses and cannot be used to make the room count indicators more comparable. We are unable to identify where the difference between the DHS and the census stems from since the census also asks for the number of rooms used for sleeping. However, we tentatively conclude that the census and LSMS enumerators might have had a similar understanding about the room count, highlighting that even the measurement of countable items can be reported according to different perceptions.

Concerning the source of drinking water, the three datasets present fairly comparable questions and answer categories (Figure 1). Roughly half of the population relies on boreholes as their main source of drinking water with the 2008 census reporting the lowest share (48%). According to the LSMS the share of households relying on boreholes is as high as 58%, which amounts to a difference of 6.7 (9.7) percentage points relative to the DHS (2008 census) data. According to the DHS more households rely on public pipes, open wells, and protected wells. The census gives a more prominent role to open wells. The remaining categories of water sources – piped into yard/dwelling, river, and other sources – are rather similar across the datasets, yet the differences found are statistically significant. As to the magnitudes, the difference of 6.7–9.7 percentage points in the prevalence of boreholes and of 4.1–5.1 percentage points in the prevalence of open wells are meaningful if they are not the result of categorization errors and if the quality of water from the two sources differs greatly. These findings suggest that for large-scale surveys to be population representative, a high number of detailed answer categories may lead to classification error. It is advisable to have general and very distinct answer categories. One also has to keep in mind that the main source of drinking water can change over the seasons. Given that the surveys were carried out at different times across the year the differences in findings are not surprising. The largest difference pertaining to drinking water is found in the time needed to fetch water (Table 2). We only have this information for the LSMS and DHS. In the DHS 28.54 minutes is recorded, which is almost twice the average time. However, this difference is artificial as the DHS asks for the total time needed, whereas the LSMS asks for the time needed one-way.

Sanitation levels influence the likelihood of household members getting certain diseases (Smith et al, 2005; WHO, 2010). All datasets clearly show that more than 80% of the Malawian population uses latrines (Figure 1, middle panel). Differences come when considering the types of latrines used. The DHS focuses on the existence of a lap because the survey is very concerned with the hygienic aspects of toilets, whereas the LSMS differentiates between latrines with and without roofs, where roofing coding indicates economic well-being. Thus, it is apparent that each data collection is subject to a different agenda and survey purpose. In contrast, the 2008 census does not introduce any differentiation of latrines. As for latrines, we find a similarly high level of consensus across datasets concerning the use of flush toilets, corresponding to approximately 3%, and the share of households that do not have a toilet, corresponding to approximately 10%.

The DHS and LSMS also contain questions about whether toilet facilities are shared with other households (Table 2). The DHS suggests a 6.7% higher fraction of households that share toilet facilities. While the DHS variable is the result of a simple 'yes or no' answer, the LSMS variable results from reading out two possible answer categories. The lower reporting of toilet sharing might result from the fact that individuals are more inclined to support the first answer possible read out to them ('no sharing') while not listening to other possible answers, which in this case pertains to toilet sharing. Moreover, questions about personal hygiene might be perceived as sensitive and response bias might arise.

There are three other key variables to describe the housing and living conditions of poor households. First, the flooring material of the main dwelling where the DHS offers different answer categories compared to the other two data collections (Figure 1, bottom panel). What is referred to as 'sand' in the DHS seems to be labeled 'mud' in the census and the LSMS, showing that analysts do not only have to know the questionnaire well but also the definition and meaning of the answer categories. While the LSMS reports a statistically significant 1.2 (3.1) percentage points more households with cement flooring compared to the DHS (2008 census), the magnitude of is difference is not meaningful since all datasets suggest that slightly more than one fifth of the households have cement floors. Second, the three datasets show similar patterns of the reported use of firewood (Figure 2, top panel) with more than 80% of the households relying on wood as fuel for cooking. As the LSMS tries to assess the overall economic conditions, it further distinguishes between collected and purchased wood demonstrating that almost all the wood used for cooking is collected.

Last, the assets owned by the households. The quantity and types of asset holdings are of interest to policymakers as fluctuations in assets have important implications for the well-being of households and can function as coping mechanisms. Moreover, in the absence of income and expenditure data, asset information can be used to construct wealth and inequality indices (McKenzie, 2005). In Table 2, bottom panel, we report about land and livestock holdings based on data from the LSMS and DHS alone. Roughly 80% of the households own land indicating the importance of agriculture for the Malawian society. At this very high level of land ownership a difference of 4.1 percentage points across surveys is negligible. However, reported livestock holdings differ by 16.7 percentage points which is statistically and economically important because it suggests that almost one fifth of the population is classified as not owning livestock according to LSMS but as livestock owners according to DHS. Such a large difference in a classification has the potential to considerably skew relevant policy decisions. Where does this difference stem from? Again, it can be explained in the way the two questions were formulated. The LSMS has a reference period of 12 months for livestock but does not include poultry. The DHS asks about the current ownership of livestock, herds, other farm livestock, and poultry. Including poultry results in a higher reported ownership of livestock. According to Gondwe and Wollny (2007) poultry farming is very popular in Malawi and neglecting its existence can have adverse implications for the policy prescriptions provided to the world’s poorest country.

In Figure 2, bottom panel, we further show a graphical representation of the prevalence of seven other assets. The questions were similarly phrased across data collections. We observe that the 2008 census and the DHS tend to report higher shares of households being in possession of the various assets. The differences between the DHS and LSMS are most pronounced for radios (7.6 percentage points) and bikes (5.2 percentage points). The difference between the census and LSMS is 18.5 percentage points for radios and 6.4 percentage points for bikes. From the survey manuals it is not evident whether possession distinguishes between functioning and non-functioning devices. The major difference in terms of survey set-up is that the census and the DHS only collect information on possessions (‘yes or no’ answer), whereas the LSMS collects detailed information on possessions (number of items, age, and the current value) of as many as 32 durables. Preceding the section on the possession of durables in the LSMS there is a long module on food and non-food expenditures. Thus, it might well be fatigue that leads to the under-reporting of these seven assets in the LSMS, which could hint at a quality-quantity trade-off in retrieving correct information even about countable items. But it might also be that due to the more explicit focus on the value of the items in the LSMS, households did not report the possession of items that lost their value.

In concluding this section, we observe that most of the detected differences across datasets are statistically significant, which is a consequence of the large sample sizes. The economic difference in the responses is by and large rather moderate, even when we triangulate the surveys statistics with those from the 2008 census data. Many of the seemingly important differences in average statistics across surveys can be attributed to a difference in the wording of the survey question. However, we find some differences in the ownership of key assets that do not seem to be the result of differences in the framing of the survey questions. We highlight that even the possession of countable goods can be documented erroneously. Nevertheless, the univariate analysis shows that overall the LSMS and DHS represent similar features of the underlying population, suggesting that each of the two datasets in themselves is a fair representation of the demographic characteristics and household infrastructure of Malawi in 2010 and representative according to the 2008 census.

5. Multivariate analysis: Linking the socio-demographic characteristics with health and education

5.1 Child nutritional status

The nutritional status of children and its correlates are widely studied to identify bottlenecks for child development and the impacts of programme interventions (Manley et al, 2013; Fernald et al, 2009; Gertler, 2004; Duflo, 2003). Analyses on children's health status and health status' determinants often use DHS and LSMS data (Yarnoff, 2011; Gomes Victora et al, 2010; Garg and Morduch, 1998). Therefore, we similarly employ the LSMS and DHS data in a correlation analysis studying the socio-demographic characteristics associated with child health. For this analysis we cannot make use of the census data as it does not contain child anthropometrics.

In line with many existing studies we assess child weight- and height-for-age (WAZ and HAZ) by making use of Z-scores expressed in terms of standard deviations from a well-nourished reference population. We employ the 2006 growth standards for attained weight and height in both datasets (WHO and UNICEF, 2009). Descriptive statistics are presented in Table 3, panel A. Although DHS is the larger survey, 2,000 fewer children were measured and weighted. Based on the survey manual we cannot deduce why the DHS has fewer observations of child anthropometrics. We would expect more observations for the DHS as all children aged 0 to 5 years were eligible whereas for the LSMS only children between six months and 5 years were considered. This problem was identified by Schoumaker (2011) who argued that the de facto omission of newborns from the measurement was due to the very detailed health module in the DHS which would not necessarily applies to newborns.

According to the DHS descriptive statistics, the Malawian children are considerably worse off compared to the LSMS (Table 3). They are more underweight and stunted with the difference being highly significant statistically and in actual magnitude. The open question is whether the DHS data, or a part of it, may over-report the severity of malnutrition due to selection bias. Keeping in mind that Malawi is the world's poorest country, the reliability of child health statistics is a key concern.

Both child samples are roughly gender-balanced and share similar features. Children are on average about 2.5 years old, although they are 3.2 months younger in the DHS sample, which is a result of the eligibility rules. Not all children were fully measured explaining why there are differences in the reported observations between age, gender and the anthropometric scores.

In the multivariate analysis we combine the child information with the household demographic and asset indicators, which further reduces the observed samples to our estimation sample. The latter has full information across all indicators. In Table 3, panel B, we assess the difference-in-means between the full survey data and the smaller estimation samples. Differences between the full and the estimation samples are small, and in many cases not statistically significant, suggesting that estimation subsamples are still a fairly good representation of the original sample.

In our effort to identify the correlates of child health, we estimate the following model for every child *i*:

$$health_i = \beta_0 + \beta_1\, age_i + \beta_2\, gender_i + \beta_3\, mat\_educ_i + \beta_4\, hhsize_i + \beta_5\, assets_i + \lambda + \varepsilon_i \qquad (1)$$

where $health_i$ is either the weight- or height-for-age Z-score of child *i*. The control variables include the age and gender of the child. Maternal education is denoted by mat_educ_i and is split in two dummy variables for primary and secondary education with no education forming the excluded category (Chen and Li, 2009; Schultz, 2002). Educational attainment is measured differently across the two surveys. From the available schooling information we construct these two simple dummy variables to ensure similarity and thus direct comparability. We further control for household size ($hhsize_i$), household assets and living conditions ($assets_i$) including mobile phone, TV, radio, land and livestock ownership. Living conditions are reflected by shared toilet facilities, the time needed to fetch water, the number of rooms, having a cement floor and access to electricity. All specifications include cluster fixed-effects, λ, to control for neighborhood level infrastructure. Standard errors are clustered at that level (Cameron and Miller, 2015) and survey weights are applied. In addition, we employ an econometrically more rigorous specification with household fixed-effects. This specification only allows us to identify the individual level correlates of child anthropometrics, namely age_i, $gender_i$ and mat_educ_i. All results are presented in Table 4.

Across surveys and specifications we find a negative relationship between age and the Z-scores, which is coherent with other studies on child anthropometrics (Rieger and Wagner, 2015). The magnitudes of the coefficient estimates are comparable across the two surveys. A negative gender effect for male children is often reported but it tends to be small and does not necessarily show up significantly (Pongou et al, 2006). Our findings are similar since the coefficient associated with male gender is negative across specifications, albeit not always statistically significant (compare WAZ model for DHS with that for LSMS, Table 4). While maternal education has been repeatedly identified as having a positive impact on child health we cannot single out this effect. This is in line with Desai and Alva (1998) who argue that education acts as a proxy for the socioeconomic status of the household.

We now turn to assets and living conditions and their correlation with child health. Here, the two datasets identify different patterns. Access to information is identified as having a positive relationship with child health. However, according to the LSMS data access to information is through mobile phone ownership (Table 4, columns 4 and 5) and whereas this is through access to television according to the DHS data (Table 4, columns 7 and 10). Moreover, the LSMS data suggest that short-term underweight is negatively correlated with sharing a toilet and positively with the number of rooms a household has. Thus, the considerable average variation across datasets in the variable ‘toilet is shared’ (6.3% difference) manifests itself in different outcomes in the multivariate analysis. Similarly, the different definitions used to count the number of rooms have bearing when relating the room count with other variables. Also, the LSMS data suggest that children living in houses with cement floor are better off (Table 4, columns 1 and 4) but this is not supported by the DHS data. The DHS data, in turn, indicate that livestock ownership is positively associated with child health (Table 4, columns 7 and 10). As identified in the previous section, the DHS also counts poultry as part of livestock suggesting important nutrition and income gains associated with poultry ownership that cannot be captured by the more narrow LSMS definition of livestock. Given the importance of poultry farming in Malawi (Gondwe and Wollny, 2007), this is an important difference highlighting the need for contextual knowledge when assessing existing data.

If the admittedly ad-hoc analysis at hand were used for policy recommendations, the correlation between the asset indicators and child health, based on LSMS and DHS data respectively, would yield different conclusions when trying to establish priority areas. Based on the results using the LSMS data one would advocate for the extension of mobile phone services and possibly for their use to disseminate health information in order to improve child health. Increased efforts for improving sanitation facilities and better housing infrastructure in general is also supported by the LSMS data. Conclusions drawn from the DHS results would identify other priority areas, namely the dissemination of (health-related) information through television and support for livestock owning households, as investments in these two areas seem to be most promising for child health.

Although the multivariate analysis of the asset indicators from the two surveys leads to different conclusions about the determinants of child health, it is reassuring that a simple additive asset index is equally positively related with children’s well-being across surveys and anthropometric indicators (Table 4, columns 2, 5, 8 and 11). This suggests that composite indicators may be more credible when aiming at drawing conclusions for the reference population. It is in line with McKenzie (2005), who demonstrated that in the absence of information on household income/expenditures, a composite wealth indicator of household infrastructure and assets is well suited to measure household inequality. Yet, these composite indicators do not lend themselves for defining priority areas for interventions and public policy.

Lastly, the individual level dynamics detected in the specification with cluster fixed-effects are supported by a more rigorous specification with household fixed-effects (Table 4, columns 3, 6, 9 and 12). In the latter specification we cannot include the household level control variables (i.e. asset indicators). They drop out for reasons of perfect multi-collinearity. But we can identify the coefficients associated with the individual level covariates. The results suggest that both surveys allow us to draw fairly similar conclusions when based on a rigorous econometric specification that only identifies individual level covariates such as age, gender and maternal education, but partials out household-level differences. Again, age is the correlate that is coherently established as being negatively related to child anthropometric status. Maternal education cannot be unanimously identified as a correlate of child health.

5.2 Level of schooling of children and young adults

We carry out the same exercise for education as we did for child health. We focus on schooling of children and young adults between 5 and 24 years, as schooling is also frequently analyzed to assess the development of human capital in a country (Akbulut-Yuksel and Turan, 2013; Picard and Wolff, 2010). In the LSMS dataset, the variable we consider is the ‘highest level of class attended’ by an individual. A similar variable from the DHS dataset is the one capturing ‘education in single years’. While the construction of the child health indicators was identical across datasets, the schooling variables differ because different questions were used across the surveys to capture schooling.

For schooling we have considerably larger datasets. The DHS dataset is more than twice as big as the LSMS dataset, which is in line with our expectation as the former surveyed more than twice as many households. According to the LSMS dataset, level 4 is the highest class level the individuals aged 5 to 24 years attend. The education in single years (from DHS) amounts to 3.5

years on average and reflects the difference in measuring schooling across the two surveys. However, the surveys represent very similar populations according to the gender and age profiles, almost 50% of the individuals are male and the average age is almost 13 years. Pronounced differences across surveys show up when considering the position of the individual in the household. More than 80% of the individuals in the LSMS dataset are children of the household head. This share is substantially lower in the DHS data with the difference amounting to 22.9 percentage points. For grandchildren the picture is reversed with only 3% of the LSMS sample representing grandchildren compared to 13.8% in the DHS. This suggests that, while the two surveys represent similar age-cohorts, the individuals of school age are found in different types of households, which is likely based on differing definitions across the two surveys.[1]

When we combine the schooling information with the household demographic and asset variables, we lose some observations. Comparing the full sample with the smaller estimation sample we detected some differences, but they are small and most of them are not statistically significant indicating that the subsamples used for the estimation are credible representations of the full datasets. In identifying the covariates of schooling we estimate a model that is equivalent to the one presented in equation (1):

$$schooling_i = \alpha_0 + \alpha_1\, age_i + \alpha_2\, gender_i + \alpha_3\, hh\_demo_i + \alpha_4\, assets_i + \lambda + \varepsilon_i \qquad (2)$$

where the outcome, $schooling_i$, is the level of schooling of individual i. The control variables include the age and gender of the individual and a vector of household demographics (hh_demo_i), which includes household size, and whether the individual is a child or grandchild of the household head. The vector $assets_i$ includes the same assets as in the child health specification. We similarly control for neighborhood level effects, λ, cluster the standard errors, and apply population weights.

[1] Descriptive statistics are not shown for the sake of brevity but made available by the authors upon request.

We also estimate a specification with household fixed-effects that only identifies the individual level covariates. All results are presented in Table 5.

Due to the large number of observations we can identify the correlates of educational attainment with more precision compared to those of child health. Across specifications and samples we coherently find that older individuals have a higher level of education, which is an expected pattern for a sample of school-aged individuals. Men and boys are less likely to go to school. Children of household heads have a higher likelihood of going to school, with the outcomes from the LSMS data being almost three times higher than that of the DHS data. The same holds for grandchildren of household heads. The findings are yet another piece of evidence against the unitary household model (Alderman et al, 1995) demonstrating that the relative position within the household matters for access to education. While it seems that household size is positively associated with education this result is reversed once we control for assets and living conditions. This feature appears in both the LSMS and the DHS data. When augmenting the specification with the asset variables, the coefficients associated with age, gender and relation to the household head only change very little and statistical significance remains, which indicates the robustness of the specifications.

The asset variables in both datasets show that information in the form of mobile phones and TV has a positive impact on educational attainment. However, the relationship between education and radio ownership differs; while it is positive and statistically significant for the LSMS, it is negative and insignificant for the DHS. This difference in correlations across datasets is also reflecting the difference in the number of radios found in the descriptive analysis (see Figure 2). The variables ‘shared toilet’ and ‘time to fetch water’ display a negative and statistically significant relationship only in the DHS data where sharing a toilet explains as much as 6% of the standard deviation in educational attainment. The time needed to fetch water explains little in economic terms. Again, these two variables had already displayed differences in the descriptive analysis stemming from the wording of the question. Interestingly, despite being differently measured, the number of rooms shows a similar relationship with education across surveys. When again employing a simple, additive wealth index, we find a positive correlation with educational attainment across surveys. The same holds for livestock and access to electricity. Thus, although the *outcome variables* are differently measured, we find only moderate differences in the correlations with the household assets and draw similar conclusions from the different datasets, which highlights that is an important determinant for proper inference. This is an important difference between the child health and the education specifications because those for education are estimated based on more than five times larger datasets.

Similar to the child health results, we also find for schooling that a specification with household fixed-effects displays coherent results across surveys for the individual-level covariates (Table 5, columns 4 and 8). Thus, across surveys individual-level dynamics are comparable. The result cautions against ad-hoc, simplistic empirical models and resulting policy conclusions. At the same time, this finding gives further support for the trend in applied microeconomics to employ carefully designed and highly data-demanding fixed-effects models.

6. Incomparability and divergences across surveys

Despite having found consistent similarities across the two datasets, we cannot argue that micro-data are a reliable source for predicting poverty and consumption trends over the years. Trend analyses are often impeded since the surveys tend to take place infrequently, i.e. generally twice within a decade. For example, Fox and Pimhidzai (2013) present evidence that lack of consistency is particularly problematic when it comes to employment data for sub-Sahara Africa. Deficits in reliability result from variations both within and across surveys. Guarcello et al. (2010) further emphasize that differences in unobservable survey characteristics, i.e. interview methods and the familiarity of the interviewers with the concept under study, lead to different estimates and ensuing conclusions. Thus, time and differences in concepts and definitions have the potential to invalidate long-term country and cross-country analyses.

Moreover, arbitrary changes to variable definitions within DHS and LSMS can seriously impair the ability to track changes over time. Comparability within and across surveys disappear entirely if different definitions are applied over time or across surveys (Srinivasan, 1994, compare our variable “Time needed to fetch water”). Most importantly, the comparability between DHS and LSMS is limited to general categories that can often only be expressed as dummy variables and tend to be of limited information content. Such categories include for example the ownership of land and livestock (Table 2). However, once we aim for analyzing cropping patterns, soil quality and irrigation methods as well as breeding success and livestock diseases we have to rely exclusively on LSMS data and can no longer employ a comparative analysis.

Similarly, when it comes to the construction of poverty profiles the two surveys exhibit considerable differences. The DHS wealth index is a composite measure of a household's cumulative living standard. The wealth index is calculated using easy-to-collect data on a household's ownership of selected assets resulting in five equally sized wealth groups. The LSMS measure of welfare used in the poverty analysis is the total annual per capita consumption reported by a household, resulting in a sample split between poor and non-poor household according to the national poverty line. Ensuing comparisons of poverty profiles have to take into account that the poverty measures of both surveys have completely different origins.

The presented flaws of micro-data and sources of inconsistency have called critics on stage. They suggest that analysts employing developing country data resort to national accounts information. Yet, Jerven (2013) shows that sub-Saharan African national income accounts are often of very poor quality since published figures tend to require a great deal of back-of-the-envelope calculations or even guesswork. Jerven (2013) further argues that census and demographic statistics tend to be more accurate since they build on DHS. Thus, based on our findings we suggests that it might rather be the national accounts data that are problematic in particular when multiple micro-data sources agree with each other but do not line up with the aggregate data.

7. Conclusions and policy implications

The last few decades have witnessed an exponential increase in the collection of micro-economic data, which will most likely continue. The availability of these data has allowed researchers to study a large array of socio-economic issues and to evaluate the effectiveness of programmes designed to solve those issues. Even though sample size calculations, sampling procedures, and ways of collecting household data have been refined in the past years, resulting in theoretically representative datasets, it is not possible to completely rule out sources of error and bias.

This paper, taking advantage of the fact that a census was carried out in Malawi in 2008, followed by two large-scale data collections in 2010, assessed the *empirical* comparability of these datasets. It concluded that descriptive statistics of most of the socio-economic variables are comparable across the surveys. Most importantly, the LSMS and DHS data are not only highly comparable but also representative as demonstrated by the comparison with the 2008 census. However, in some cases, seemingly identical variables displayed different dynamics. We demonstrated that these differences are driven by definitional variations and the way the survey questions were formulated. By differentiating answer categories and giving different reference points, the datasets do not display exactly the same information despite being carried out by the same institution.

The multivariate analysis has shown that researchers should be careful when claiming representativeness as a considerable number of observations gets lost due to missing information across variables. Whenever multivariate analyses are *de facto* subsample analyses, such as our child health estimates, it is even more likely that different datasets yield different conclusions. We demonstrated that LSMS and DHS point to different correlates of welfare outcomes, and thus point to different directions for further investigation.

In poverty-stricken countries such as Malawi this might have substantial implications on the identification of priority intervention areas. Despite loss of representativeness we draw more coherent conclusions across datasets when implementing demanding empirical specifications that employ fixed-effects at the lowest possible level. However, these specifications tend not to be suited for drawing policy conclusions since many interventions such as infrastructure development, health service expansion or decentralization efforts are implemented at community level, sometimes at household level but not at the individual level. Our findings suggest that the poorer the country is, the more funds should be invested in the collection of data to ensure that the evidence derived for policy making is properly identifying the areas of highest potential impact. Larger surveys allow more coherent representation of the underlying survey population especially when multivariate and subsample analyses are employed. Our findings further call for *empirical* validation studies and comparative approaches, especially if multivariate analyses are employed and sensitive conclusions are drawn (Epple et al, 2015).

When turning to divergences across surveys we briefly discussed that there are multiple dimensions in the DHS and LSMS datasets that cannot be compared and that comparability hinges largely on similar sampling frames and definitions of the variables being compared. Comparability within and across surveys disappears entirely if different definitions are applied over time or across surveys. Overall, this paper shows that micro-data from accredited sources are rich and credible fonts of information that can be used to study possible paths for human development across the globe. When working with existing data sources, researchers are advised to acquire profound contextual knowledge, to familiarize themselves with the peculiarities of the chosen survey and corresponding dataset, and to take all possible data-related limitations into account before starting the analysis. If empirical analyses respect the data requirements laid out by theory, micro-data from authoritative sources constitute reliable grounds for factually informed policy recommendations.

References

Akbulut-Yuksel, M. and Belgi, T. (2013) Left behind: intergenerational transmission of human capital in the midst of HIV/AIDS. Journal of Population Economics 26(4): 1523–1547.

Alderman, H., Chiappori, P.A., Haddad, L., Hoddinott, J. and Kanbur, R. (1995) Unitary versus collective models of the household: Is it time to shift the burden of proof? World Bank Research Observer 10(1): 1–19.

Alwin, D.F. (1989) Problems in the estimation and interpretation of the reliability of survey data. Quality and Quantity 23: 277–331.

Bound, J., Brown, C. and Mathiowetz, N. (2001) Measurement error in survey data. In: J.J. Heckman and E. Leamer (eds.) Handbook of Econometrics Volume 5: 3705–3843. Amsterdam, The Netherlands: Elsevier.

Brick, J.M. and Kalton, G. (1996) Handling missing data in survey research. Statistical methods in medical research 5(3): 215–238.

Cameron, A.C. and Miller, D.L. (2015) A practitioner's guide to cluster-robust inference. Journal of Human Resources 50(2): 317-372.

Checkley, W., Gilman, R.H., Black, R.E., Epstein, L.D., Cabrera, L., Sterling, C.R. and Moulton, L.H. (2004) Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. The Lancet 363(9403): 112–118.

Chen, Y. and Hongbin L. (2009) Mother's education and child health: Is there a nurturing effect? Journal of Health Economics 28(2): 413–426.

Deaton, A. (1997) The analysis of household surveys: a microeconometric approach to development policy. World Bank Publications.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world. Review of Economics and statistics, 87(1), 1-19.

Desai, S. and Alva, S. (1998) Maternal education and child health: Is there a strong causal relationship? Demography 35(1): 71–81.

Duflo, E. 2003. Grandmothers and granddaughters: Old-age pensions and intrahousehold allocation in South Africa. World Bank Economic Review 17(1): 1–25.

Elbers, C., Lanjouw, J.O. and Lanjouw, P. (2003) Micro-level estimation of poverty and inequality. Econometrica 71(1): 355–364.

Epple, D., Romano, R.E. and Urquiola, M. (2015) School vouchers: A survey of the economics literature. National Bureau of Economic Research, Working Paper Series no. 21523.

Fernald, L.C.H., Gertler, P. and Neufeld, L.M. (2009) 10-Year effect of Oportunidades, Mexico’s conditional cash transfer programme, on child growth, cognition, language, and behavior: A longitudinal follow-up study. The Lancet 374(9706): 1997–2005.

Fox, L. and Pimhidzai, O. (2013) Different Dreams, Same Bed: Collecting, Using, and Interpreting Employment Statistics in Sub-Saharan Africa -- the Case of Uganda. World Bank Policy Research Working Paper No. 6436. Garg, A. and Morduch, J. (1998) Sibling rivalry and the gender gap: Evidence from child health outcomes in Ghana. Journal of Population Economics 11(4): 471–493.

Gertler, P. (2004) Do conditional cash transfers improve child health? Evidence from PROGRESA’s controlled randomized experiment. American Economic Review 94(2): 336–341.

Gondwe, T.N. and Wollny, C.B.A. (2007) Local chicken production system in Malawi: Household flock structure, dynamics, management and health. Tropical Livestock Health and Production 39(2): 103–113.

Gomes Victora, C., de Onis, M. Hallal, P.C., Blössner, M. and Shrimpton, R. (2010) Worldwide timing of growth faltering: Revisiting implications for interventions. Pediatrics 125(3): e473–e480.

Griliches, Z. (1986) Economic data issues. In: Z. Griliches and M. Intriligator (eds.) Handbook of econometrics Volume 3: 1465–1514. Amsterdam, The Netherlands: Elsevier.

Grosh, M. E., and Glewwe, P. (1996). Household survey data from developing countries: progress and prospects. The American Economic Review, 86(2), 15-19.

Guarcello, L., Kovrova, I., Lyon, S., Manacorda, M. and Rosati, F. C. (2010) Towards Consistency in Child Labour Measurement: Assessing the Comparability of Estimates Generated by Different Survey Instruments. Understanding Children's Work Programme Working Paper.

Heckman, J.T. (2001) Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. Journal of Political Economy, 109(4): 673–748.

Holt, D. (1985) Review of planning and analysis of observational studies by William E. Cochran. Journal of the American Statistical Association 80: 772–73.

Jerven, M. (2013) Poor numbers: how we are misled by African development statistics and what to do about it. Cornell University Press.

Langkamp, D.L., Lehman, A. and Lemeshow, S. (2010) Techniques for handling missing data in secondary analyses of large surveys. Academic pediatrics 10(3): 205–210.

Little, R.J.A. and Rubin, D.B. (2002) Statistical analysis with missing data. Second edition, Hoboken, NJ: John Wiley & Sons.

Manley, J., Gitter, S. and Slavchevska, V. (2013) How effective are cash transfers at improving nutritional status? World Development 48: 133–155.

McKenzie, D.J. (2005) Measuring inequality with asset indicators. Journal of Population Economics 18(2): 229–260.

Morgenstern, O. (1963) On the accuracy of economic observations. Second edition. Princeton, NJ: Princeton University Press.

National Statistical Office and ICF Macro. (2011) Malawi demographic and health survey 2010. Zomba, Malawi, and Calverton, Maryland, USA: National Statistical Office and ICF Macro.

National Statistical Office. (2012) Integrated household survey 2010–2011; Household socio-economic characteristics (report). Zomba. Malawi.

National Statistical Office. (2008) 2008 Population and housing census (preliminary report). Zomba. Malawi.

Orcutt, G.H. (1962) Microanalytic models of the United States economy: need and development. The American Economic Review 52(2): 229–240.

Picard, N. and Wolff, F.C. (2010) Measuring educational inequalities: a method and an application to Albania. Journal of Population Economics 23(3): 989–1023.

Pongou, R., Ezzati, M. and Salomon, J.A. (2006) Household and community socioeconomic and environmental determinants of child nutritional status in Cameroon. BMC Public Health 6:98.

Randolph, T.F., Schelling, E., Grace, D., Nicholson, C.F., Leroy, J.L., Cole, D.C., Demment, M. W., Omore, A., Zinsstag, J. and Ruel, M. (2007) Role of livestock in human nutrition and health for poverty reduction in developing countries. Journal of livestock science 85(11): 2788–2800.

Ravallion, M., and Chen, S. (1997) What can new survey data tell us about recent changes in distribution and poverty? World Bank Economic Review 11(2): 357–382.

Ravallion, M. (2003) The debate on globalization, poverty and inequality: why measurement matters. International Affairs 79(4): 739–753.

Rieger, M. and Wagner, N. (2015). Child health, its dynamic interaction with nutrition and health memory – Evidence from Senegal. Economics & Human Biology 16: 135–145.

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

Schoumaker, B. (2011) Omissions of births in DHS birth histories in sub-Saharan Africa: Measurement and determinants. Proceedings of the 2011 Annual Meeting of the Population Association of America; 31 March–2 April, Washington D.C.

Schultz, T.P. (2002) Why governments should invest more to educate girls. World Development 13: 827–846.

Smith, L.C., Ruel, M.T., and Ndiaye, A. (2005) Why is child malnutrition lower in urban than in rural areas? Evidence from 36 developing countries. World Development 33(8): 1285–1305.

Srinivasan, T.N. (1994) Data base for development analysis: An overview. Journal of Development Economics 44 (1), 3–27.

Sutherland, H., Taylor, R., and Gomulka, J. (2002) Combining household income and expenditure data in policy simulations. Review of Income and Wealth 48(4): 517–536.

WHO. (2010) World health statistics 2010. Geneva, Switzerland: World Health Organization.

WHO and UNICEF. (2009) WHO child growth standards and the identification of severe acute malnutrition in infants and children. A joint statement by the World Health Organization and the United Nations Children's Fund. Geneva, Switzerland: WHO Press. Available at <http://www.who.int/nutrition/publications/severemalnutrition/9789241598163_eng.pdf>.

World Bank. (2015) World Development Indicators -- GDP per capita (current US$). World Bank national accounts data. Washington, D.C.: World Bank.

Yarnoff, B. (2011) Household allocation decisions and child health: can behavioral responses to vitamin A supplementation programmes explain heterogeneous effects? Journal of Population Economics 24(2): 657–680.

Tables and Figures

Table 1: Total number of observations and demographic composition

	2008 census		**LSMS**		**DHS**		Difference in the share Census-LSMS	Difference in the share Census-DHS	Difference in the share LSMS-DHS
	Obs.	Share (%)	Obs.	Share (%)	Obs.	Share (%)			
Total number of households	2,892,913	-	12,271	-	27,307	1	-	-	-
Male individuals	6,358,933	48.63	27,560	48.86	56,874	48.71	-0.23	-0.09	0.14
Female individuals	6,718,227	51.37	28,849	51.14	59,875	51.29	0.23	0.09	-0.14
Age cohorts									
0‒4	2,369,928	18.12	11,162	19.79	19,974	17.11	-1.66	1.01	2.68
5‒14	3,638,548	27.82	17,116	30.34	37,468	32.09	-2.52	-4.27	-1.75
15‒24	2,516,949	19.25	10,571	18.74	19,939	17.08	0.51	2.17	1.66
25‒34	1,930,447	14.76	8,112	14.38	14,673	12.57	0.38	2.19	1.81
35‒44	1,064,509	8.14	4,749	8.42	9,086	7.78	-0.28	0.36	0.64
45‒54	612,808	4.69	2,845	5.04	6,304	5.40	-0.36	-0.71	-0.36
>=55	943,485	7.22	3,860	6.84	9,305	7.97	0.37	-0.76	-1.13

Table 2: Household head and housing characteristics, and land and livestock ownership

	Population adjusted descriptive statistics						Difference in means	*p*-value of difference in means
	LSMS			**DHS**				
	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.		
Characteristics of the household head								
Age	12,268	42.263	16.404	24,825	42.716	16.313	-0.453	0.012
Male	12,268	0.762	0.426	24,825	0.719	0.449	0.043	0.000
Housing characteristics								
Electricity in the household	12,271	0.071	0.258	24,785	0.087	0.281	-0.015	0.000
Number of habitable rooms	12,271	2.546	1.178	24,758	1.919	0.923	0.627	0.000
Time needed to fetch water	12,254	14.659	59.388	24,655	28.544	31.052	-13.884	0.000
Toilet is shared with other households	11,151	0.366	0.482	21,764	0.432	0.495	-0.067	0.000
Land and livestock								
Land	12,271	0.834	0.372	24,818	0.793	0.405	0.041	0.000
Livestock	12,271	0.430	0.495	24,819	0.598	0.490	-0.167	0.000

Note: The 2008 census only contains information about the gender of the household head, the number of rooms, and electricity. The share of male-headed households is 0.729, the average number of rooms is 1.866 and the share of households with electricity is 0.072.

Table 3: Child anthropometrics in the LSMS and the DHS datasets

Panel A	Population adjusted descriptive statistics						DiM: LSMS-DHS	*p*-value of DiM		
	LSMS			DHS						
	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.				
WAZ	7,263	-0.468	1.134	4,895	-0.806	1.124	0.337	0.000		
HAZ	7,115	-1.383	1.636	4,895	-1.774	1.612	0.391	0.000		
Child is male	7,539	0.507	0.500	5,611	0.487	0.500	0.020	0.025		
Age in months	7,497	32.433	17.112	5,611	29.228	17.121	3.205	0.000		
Fully measured	7,539	0.974	0.160	5,611	0.932	0.251	0.041	0.000		
Panel B	Sub-sample employed in the multivariate analysis						DiM: LSMS full and estimation sample	*p*-value of DiM	DiM: DHS full and estimation sample	*p*-value of DiM
	LSMS			DHS						
	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.				
WAZ	5,264	-0.405	1.133	4,045	-0.801	1.114	-0.063	0.002	-0.005	0.834
HAZ	5,148	-1.347	1.673	4,045	-1.768	1.594	-0.035	0.244	-0.006	0.860
Child is male	5,264	0.507	0.500	4,045	0.488	0.500	-0.001	0.929	-0.001	0.923
Age in months	5,264	31.960	15.173	4,045	29.468	16.725	0.473	0.108	-0.240	0.493
Fully measured	5,264	0.994	0.076	4,045	1	0	-0.021	0.000	-0.068	0.000

Note: DiM abbreviates difference-in-means. WAZ (HAZ) refers to the weight-for-age Z score (height-for-age Z score) indicator.

Table 4: Child health regressions for the LSMS and the DHS dataset

	LSMS						DHS					
	WAZ			HAZ			WAZ			HAZ		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Age in months	-0.016***	-0.016***	-0.012***	-0.009***	-0.009***	-0.009***	-0.011***	-0.011***	-0.009***	-0.019***	-0.019***	-0.015***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Child is male	-0.085**	-0.085**	-0.137*	-0.213***	-0.209***	-0.080	-0.056	-0.058	-0.113	-0.169**	-0.176***	-0.248**
	(0.038)	(0.038)	(0.073)	(0.058)	(0.058)	(0.124)	(0.049)	(0.049)	(0.070)	(0.067)	(0.067)	(0.101)
Mother has primary education	-0.068	-0.073	-0.625***	-0.049	-0.046	-0.980	-0.064	-0.070	0.244	-0.088	-0.101	0.279
	(0.061)	(0.061)	(0.203)	(0.091)	(0.089)	(0.628)	(0.075)	(0.076)	(0.340)	(0.133)	(0.133)	(0.521)
Mother has secondary education	0.036	0.061	-0.633*	0.021	0.059	0.026	0.009	0.022	0.458	0.005	0.008	0.708
	(0.062)	(0.060)	(0.352)	(0.125)	(0.124)	(0.686)	(0.113)	(0.113)	(0.458)	(0.169)	(0.165)	(0.702)
Household size	-0.011	-0.004		-0.022	-0.014		0.009	0.004		0.036	0.021	
	(0.014)	(0.012)		(0.018)	(0.017)		(0.015)	(0.012)		(0.022)	(0.018)	
Household owns a mobile phone	0.079	0.082		0.180**	0.151*		0.071	0.090		0.068	0.093	
	(0.052)	(0.057)		(0.078)	(0.088)		(0.062)	(0.064)		(0.086)	(0.087)	
Household owns a TV	0.080			-0.258			0.252**			0.443***		
	(0.111)			(0.175)			(0.106)			(0.136)		
Household owns a radio	0.013			-0.132**			0.002			0.071		
	(0.042)			(0.065)			(0.064)			(0.077)		
Household uses a shared toilet	-0.122**			-0.054			0.010			-0.071		
	(0.050)			(0.070)			(0.054)			(0.077)		
Time needed to fetch water	-0.000			-0.001			0.000			-0.000		
	(0.000)			(0.000)			(0.001)			(0.001)		
Number of rooms	0.036*			0.019			-0.030			-0.063		
	(0.021)			(0.030)			(0.039)			(0.053)		
Household owns land	-0.118			0.075			-0.149			-0.155		
	(0.090)			(0.131)			(0.096)			(0.118)		
Household owns livestock	-0.013			0.076			0.190**			0.189*		
	(0.053)			(0.095)			(0.075)			(0.100)		
Floor in main dwelling is made of cement	0.136**			0.278**			0.074			0.014		
	(0.069)			(0.131)			(0.090)			(0.122)		
Household has electricity	0.248			0.744***			-0.043			-0.490		
	(0.154)			(0.229)			(0.303)			(0.386)		
Wealth index		0.048***			0.030			0.042**			0.073***	
		(0.015)			(0.022)			(0.021)			(0.027)	
Observations	5,264	5,264	5,264	5,148	5,148	5,148	4,045	4,045	4,045	4,045	4,045	4,045
Number of clusters	761	761	4153	761	761	4079	819	819	2945	819	819	2945
R^2 (within)	0.069	0.063	0.073	0.025	0.016	0.024	0.048	0.041	0.055	0.063	0.058	0.064
ρ	0.265	0.264	0.606	0.309	0.306	0.595	0.276	0.271	0.568	0.269	0.264	0.534

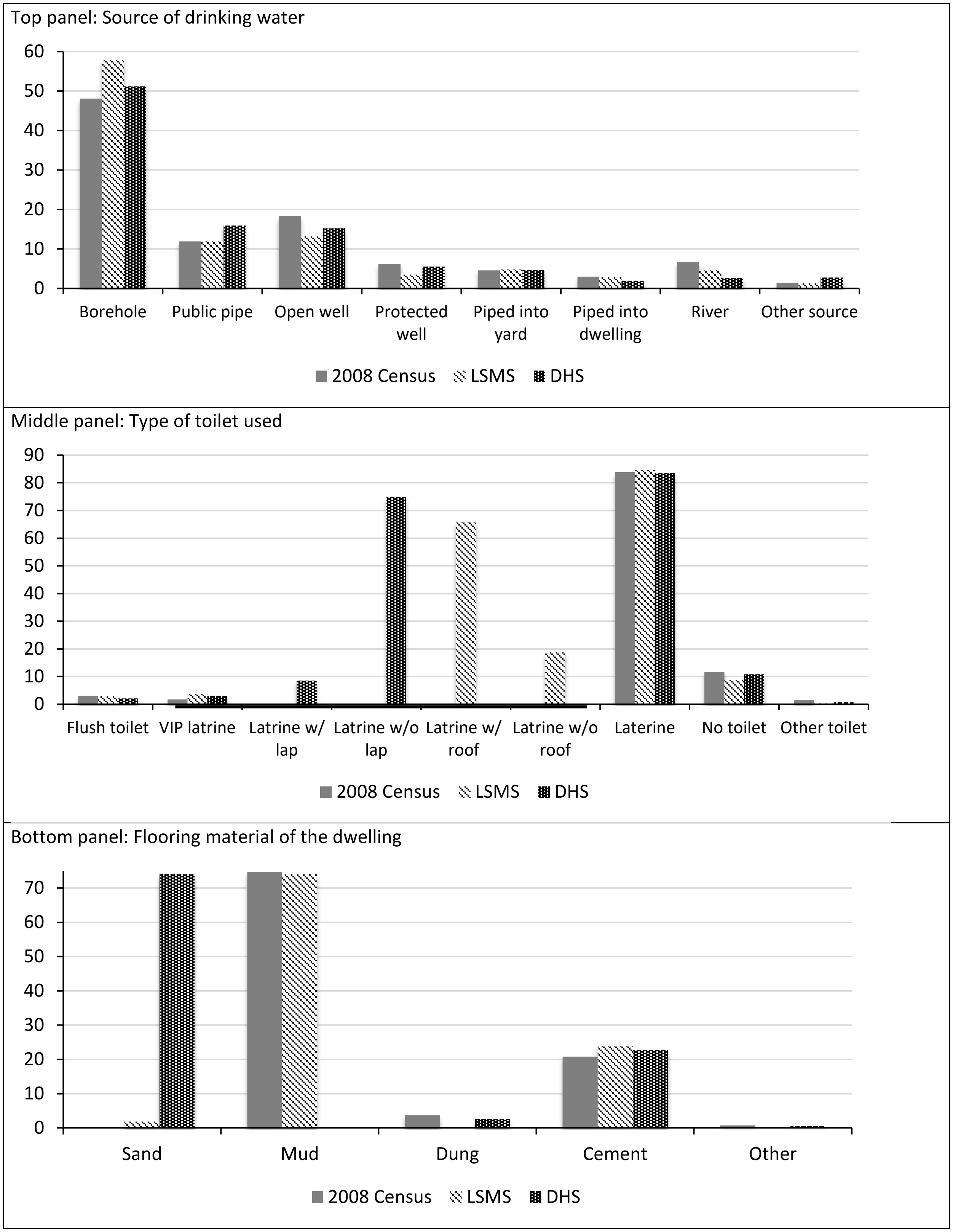
Note: Child health regressions with neighborhood and household-fixed effects. Standard errors clustered at the respective cluster-level are in parentheses. ***/**/* indicates significance at the 1/5/10 percent level, respectively.

Table 5: Schooling levels for 5–24 years old in the LSMS and the DHS datasets

	LSMS: Highest class level attended				DHS: Education in single years			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age	0.467***	0.459***	0.460***	0.455***	0.441***	0.436***	0.438***	0.438***
	(0.007)	(0.007)	(0.007)	(0.005)	(0.006)	(0.005)	(0.006)	(0.004)
Individual is male	-0.112***	-0.114***	-0.112***	-0.148***	-0.048**	-0.066***	-0.055**	-0.111***
	(0.034)	(0.033)	(0.033)	(0.035)	(0.023)	(0.022)	(0.022)	(0.025)
Child of the household head	1.850***	1.750***	1.795***	1.658***	0.602***	0.589***	0.613***	0.639***
	(0.106)	(0.102)	(0.103)	(0.117)	(0.045)	(0.044)	(0.044)	(0.056)
Grandchild of the household head	1.715***	1.629***	1.723***	1.325***	0.495***	0.491***	0.564***	0.080
	(0.145)	(0.143)	(0.145)	(0.167)	(0.047)	(0.048)	(0.048)	(0.089)
Household size	0.026**	-0.050***	-0.034***		0.026***	-0.053***	-0.017**	
	(0.011)	(0.011)	(0.011)		(0.009)	(0.009)	(0.008)	
Household owns a mobile phone		0.554***	0.556***			0.446***	0.408***	
		(0.054)	(0.053)			(0.036)	(0.037)	
Household owns a TV		0.554***				0.505***		
		(0.098)				(0.068)		
Household owns a radio		0.104**				-0.012		
		(0.046)				(0.033)		
Household uses a shared toilet		-0.038				-0.202***		
		(0.053)				(0.030)		
Time needed to fetch water		-0.000				-0.002***		
		(0.000)				(0.001)		
Number of rooms		0.191***				0.211***		
		(0.022)				(0.023)		
Household owns land		-0.168				-0.078		
		(0.106)				(0.057)		
Household owns livestock		0.170***				0.136***		
		(0.047)				(0.035)		
Floor in main dwelling is made of cement		0.427***				0.559***		
		(0.069)				(0.052)		
Household has electricity		0.574***				1.107***		
		(0.206)				(0.170)		
Wealth index			0.232***				0.236***	
			(0.017)				(0.010)	
Observations	25,274	25,274	25,274	25,274	51,022	51,022	51,022	51,022
Number of clusters	768	768	768	9,624	849	849	849	18,984
R^2 (within)	0.537	0.563	0.560	0.604	0.562	0.589	0.584	0.613
ρ	0.226	0.137	0.154	0.610	0.197	0.118	0.130	0.583

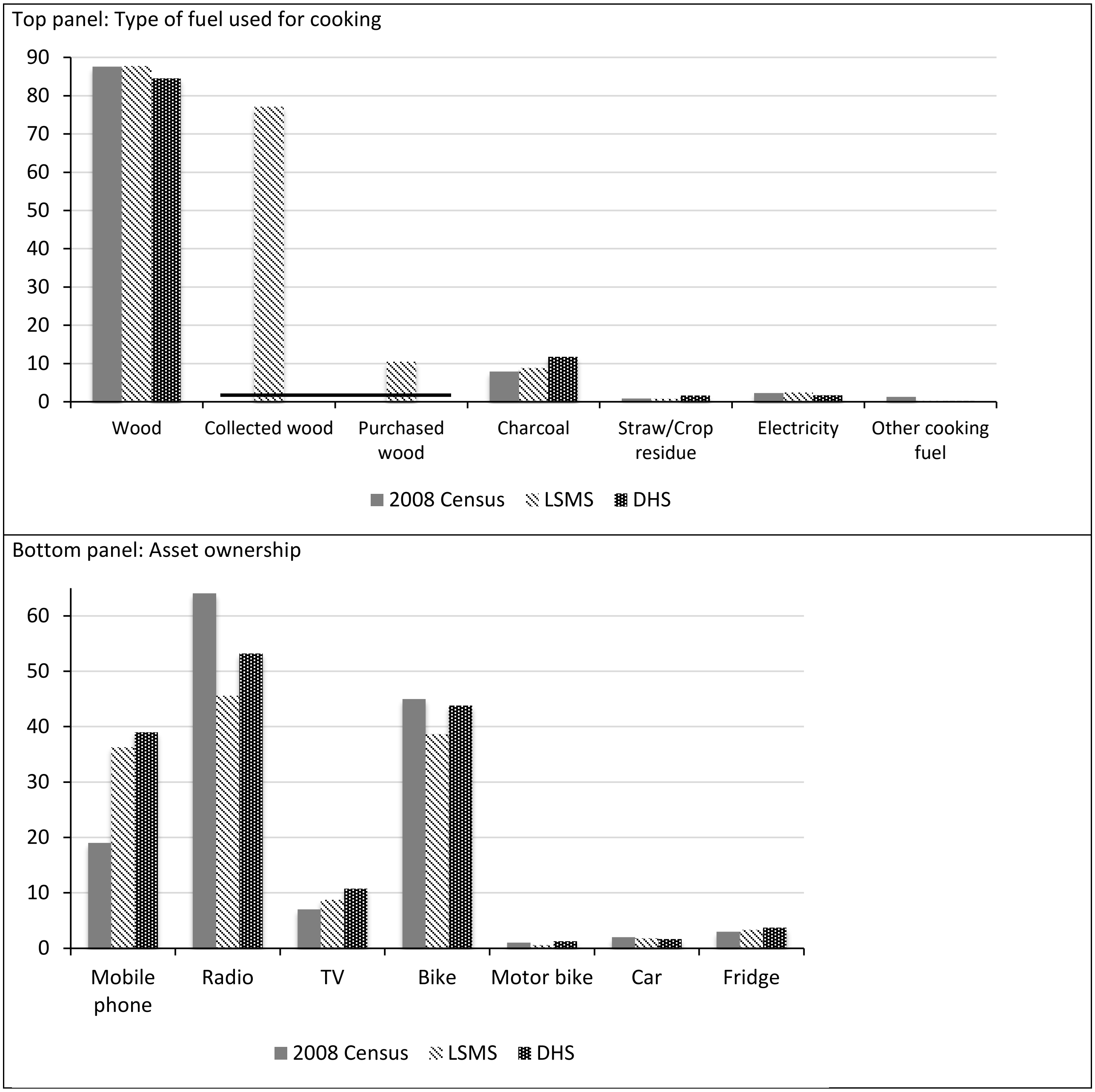
Note: Schooling regressions with neighborhood and household-fixed effects. Standard errors clustered at the respective cluster-level are in parentheses. ***/**/* indicates significance at the 1/5/10 percent level, respectively.

Figure 1: Source of drinking water, type of toilet used and flooring material of the dwelling (in %)



Note: VIP latrines are ventilated improved pit latrines. Latrines are presented as overall category and split by characteristics into latrines with/without lap and latrines with/without roof. The sub-categories add up to the overall category latrine.

Figure 2: Type of fuel used for cooking and asset ownership (in %)



Note: Wood is presented as overall category and split by type of acquisition into collected/purchased wood. The sub-categories add up to the category wood. The 2008 census only contains information on telephone ownership not on mobile phones. Note that the ownership shares are per asset group and do not add up to 100% across assets.