National Bureau of Statistics

United Republic of Tanzania

Sample Design for the National Panel Survey

Justin Sandefur

August 12, 2009

Contents

1 Overview and objectives of the survey

As part of the monitoring framework for the National Strategy for Growth and Poverty Reduction (MKUKUTA), the National Bureau of Statistics (NBS) has been commissioned to carry out an annual, nationwide, longitudinal, household survey covering both mainland Tanzania and Zanzibar. The NPS has three broad objectives: (i) to monitor progress on a range of MKUKUTA indicators, (ii) to improve understanding of poverty dynamics in Tanzania at the household level, (iii) to evaluate the impact of major development initiatives.

The NPS will comprise a multi-stage, stratified, random sample of Tanzanian households. The sample of 3,280 households will be representative of the nation as a whole and provide reliable estimates of key socio-economic

indicators for each of four strata: mainland rural areas, Dar es Salaam, other mainland urban areas, and Zanzibar.

The NPS sample will link to two other major surveys: the 2008 Agricultural Census Sample Survey (ACSS) and the 2007 Household Budget Survey (HBS). The villages drawn for the rural portion of the NPS will be a subsample of the villages surveyed by the ACSS. See NBS (**?**) for details on the ACSS sample design. Furthermore, roughly half of the NPS sample will be drawn from villages and enumeration areas covered by the 2007 HBS. A panel of approximate 1,500 households will be formed by returning to a subset of HBS households in these clusters, as described in section 4.1.

2 Population of study and primary sampling units

The unit of observation for the purposes of this document is a household. In urban areas, the primary sampling unit (PSU) for the NPS is an enumeration area (EA) from the 2002 Population and Housing Census. In rural areas, the PSUs will constitute entire villages, drawn from the population of villages recorded in the same census.[1]

2.1 Choosing the number of clusters

The objective of the NPS sample design is to produce the most reliable possible estimate of the population mean of some indicator, say household consumption, within a given budget constraint for the survey. In order to minimize costs, multilevel or clustered sampling is attractive because there are presumably high fixed costs – in terms of travel and administrative work

[1]For the purposes of the community questionnaire, interviews will refer to the entirety of the *mtaa* (or possibly plural *mitaa*) which the EA touches, despite the fact that an *mtaa* will normally incorporate several EAs. This decision acknowledges the fact that the boundaries of census EAs are not widely known by residents and are not expected to have any particular economic or social significance.

– to adding new geographic areas to the sample. The downside to clustered sampling is that it increases the size of the standard errors for any statistics produced by the sample.

Consider the following model, where y is an outcome variable of interest, i indexes households and j indexes groups (enumeration areas):

$$y_{ij} = \mu + \upsilon_j + \omega_{ij} \tag{1}$$

The error term is decomposed into a common group element, υ_j with variance τ^2, and a household specific component, ω_{ij} with variance σ^2. These parameters can be combined to yield the intraclass correlation

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

which measures "the proportion of the total population variance ($\tau^2 + \sigma^2$) across groups as opposed to within groups" (**?**, p. 4-22).[2]

One possible goal in choosing a sample design is to minimize the variance of the estimate of μ, subject to a total budget constraint for the survey, B. Using the expression for $var(\hat{\mu})$ given in Snijders (**?**) this problem can be expressed as

$$\begin{array}{rrcl} \min\limits_{n,J} & var(\hat{\mu}) & = & \dfrac{n\tau^2 + \sigma^2}{nJ} \\ & \text{s.t. } B & \geq & (n+c)J \end{array} \tag{2}$$

where c is the fixed cost of adding another group to the survey, and the unit

[2]This tradeoff between efficiency and cost is captured in the design effect:

$$\begin{array}{rcl} deff & = & \dfrac{\text{squared standard error under the chosen design}}{\text{squared standard error from a simple random sample}} \\ & = & \sqrt{1 + (n-1)\rho} \end{array}$$

where the 'simple random sample' in the denominator of the first line has the same total number of households as the chosen design. For values of $deff > 1$, the design sacrifices statistical efficiency for the sake of cost savings.

cost of adding a household within a cluster is the numeraire. Solving the problem in (2) yields an expression for the optimal group size with a given budget,

$$n^*(c, \tau^2, \sigma^2, B) = \frac{\sigma\sqrt{c}}{\tau} \tag{3}$$

Equation (3) gives an optimal cluster size which can be substituted into the budget constraint to yield the optimal number of groups.

The most recent available nationally-representative data source for household consumption in Tanzania – the 2000/01 Household Budget Survey (HBS) – produces the following estimates of the between-group and within-group standard deviations of real total household consumption per adult equivalent: $\hat{\tau} = 9,531.6$, $\hat{\sigma} = 15,785.1$, and $\rho = .267$. If we make the (somewhat arbitrary, but seemingly conservative) estimate that an additional cluster costs the equivalent of ten extra households ($c = 10$), this implies an optimal cluster size of 5.24 households.

To illustrate this calculation more concretely, Table 1 computes the standard error of mean consumption using various sample designs. The total projected cost of the survey is the same in all cases. These cost projections are again based on the assumption that $c = 10$. Standard errors are computed in two different ways. First, estimates of τ and σ from the HBS data are plugged into the equation for the variance of the mean based on a two level model as in equation (1). Second, as a check, the restrictions imposed by equation (1) are relaxed and standard errors are estimated by bootstrapping 100 clustered samples of the given size from the 2001 HBS data. The results are consistent across both approaches: the optimal sample design for the current NPS budget (and the assumption that $c = 10$) would include 5 households per cluster and 437 mainland clusters.

In practice, a cluster size of 8 households per PSU was chosen. The decision to opt for larger clusters reflects a compromise with other logistical constraints on the survey team not captured in the estimate of c (e.g., travel fatigue, advantages from building a rapport with villagers over multiple days,

etc.).

3 First stage sampling: Selecting clusters

The selection of clusters will be stratified along two dimensions. The first dimension comprises administrative the eight administrative zones of Tanzania (including Zanzibar). The second dimension consists of three strata within the mainland sample: (i) rural areas, (ii) Dar es Salaam, and (iii) other urban areas on the mainland.

The primary motivation for stratifying the sample in the first stage is to produce estimates for sub-populations (e.g., rural areas versus urban areas, or particular zones) with relatively similar confidence intervals.[3]

Formally, returning to the model of the previous section, our task is to choose the number of clusters in each of k strata,

$$J = \sum_{j=1}^{k} J_j$$

such that the variance of the estimate of the mean is the same in each stratum:

$$var(\hat{\mu}) = \frac{n\tau_a^2 + \sigma_a^2}{nJ_a} = \frac{n\tau_b^2 + \sigma_b^2}{nJ_b}$$

for all clusters a and b. Combining these two expressions yields

$$J_a = Jv_a / \sum_{j=1}^{k} J_j \tag{4}$$

where

$$v_a \equiv n\tau_a^2 + \sigma_a^2.$$

[3]An alternative motivation would be to achieve the most precise possible estimates for population parameters, taking into account heterogeneity between strata in the population. Note that goal of achieving equal confidence intervals in each strata implies more drastic re-weighting – that is, a greater departure from a simple random sample.

Table 1: Alternative Sample Designs for Same Total Survey Cost

n = no. HHs/cluster	J = no. clusters	N = total no. HHs	Standard Error of Mean Consumption: Model Prediction	Standard Error of Mean Consumption: Simulation using HBS	Standard Error of Mean Log Consumption: Model Prediction	Standard Error of Mean Log Consumption: Simulation using HBS
24	193	4,625	724.8	731.7	0.4623	0.0310
20	218	4,368	687.8	699.5	0.4606	0.0300
16	252	4,032	649.9	697.6	0.4484	0.0287
12	298	3,574	612.2	682.1	0.4402	0.0264
10	328	3,276	594.5	624.0	0.4535	0.0263
8	364	2,912	578.9	609.1	0.4530	0.0238
6	410	2,457	568.6	583.7	0.4586	0.0234
5	437	2,184	567.5*	564.7*	0.4658	0.0229
4	468	1,872	572.1	605.3	0.4517	0.0229
3	504	1,512	587.4	677.7	0.4326*	0.0228*
2	546	1,092	628.2	610.2	0.4714	0.0248
1	596	596	755.6	622.9	0.5116	0.0282

Each row corresponds to a different possible sample design. Asterisks denote the optimal sample design by each criterion. The number of clusters and cluster sizes in the first two columns are adjusted to hold the total cost of the survey constant. Cost estimates are based on the assumption that the fixed cost of an extra cluster is equivalent to the variable cost of ten households, i.e., $c = 10$. All estimates are based on the full sample of 22,176 households from the 2000/01 HBS. Consumption refers to real total monthly household consumption per adult equivalent. 'Model predictions' are calculated based on the two-level model in equation 1 of the text, using data to estimate τ and σ. The estimates from the HBS sample are computed directly by drawing a sample of the specified design from the HBS data and computing sample means.

Table 2: Sample Share by Strata to Equalize Standard Errors

		Necessary Sample Share for Equal Standard Errors				For Comparison	
	NPS Sample	Per Cap. Cons.	Clean Water	Maize Yield	Fert. Use	Pop. Share	HBS Sample
Rural	65.0	11.5	59.5	NA	NA	76.9	31.1
Dar	17.5	56.6	13.1	NA	NA	7.2	33.3
Other Urb.	17.5	31.8	27.4	NA	NA	15.9	35.6
	100.0	100.0	100.0	NA	NA	100.0	100.0
Northern	16.3	12.0	12.4	18.6	22.1	16.5	11.3
Central	6.2	5.8	13.9	4.2	1.1	8.6	5.4
S. High.	16.6	10.7	12.3	16.2	27.2	11.3	11.6
Western	14.3	24.5	16.6	7.3	8.8	19.1	9.5
Lake	14.1	13.3	14.1	26.7	2.4	19.5	11.1
Southern	19.8	12.4	19.5	14.2	32.9	9.3	8.4
Eastern	12.7	21.3	11.2	12.8	5.6	15.8	42.6
	100.0	100.0	100.0	100.0	100.0	100.0	100.0

In words, equation 4 simply stipulates that to achieve the same precision within each stratum, sample sizes should be proportional to the variance (within and between) in a given stratum. This result is applied below to calculate the optimal sample sizes in each stratum below.

3.1 Stratification: Rural areas, Dar es Salaam, and other urban areas

For Zanzibar, the NPS sample will be stratified between rural and urban areas. Equal proportions for these two strata were chosen as a compromise between the desire for greater precision in poverty estimates (implying a need for more observations in urban areas with higher inequality) and a focus on rural agriculture.

Within the mainland sample, the NPS clusters will be stratified between rural and urban areas, and within urban areas between Dar es Salaam and other urban areas. The division of the sample between rural and urban

areas reflects the desire for a minimum rural sample of approximately 2,000 households. Thus the rural stratum is arbitrarily set to comprise 65% of the total mainland sample.

The division of the urban sample between Dar es Salaam and other urban areas is calibrated to produce estimates of roughly equal precision in each stratum along various dimensions (consumption, maize yields, etc.), as described in the introduction to this section. As seen in the top panel of Table 2, focusing on household consumption suggests a need to concentrate the sample in Dar es Salaam, while focusing on access to clean water suggests concentrating on urban areas outside Dar es Salaam. As a simple compromise, the chosen sample offers an even split between Dar es Salaam and other mainland urban areas.

3.2 Stratification: Administrative zones

The mainland sample is stratified by administrative zones. The sample size for each zone is assigned according to equation 4 using the figures in Table 2. Once again, note that the desire to measure various outcome indicators pulls the sample in opposite directions. To get comparable precision on consumption in each zone would require a large sample in the Western zone for instance, while getting comparable estimates for fertilizer usage implies a very small sample in the Western zone. The compromise chosen was to take a simple average of the sample weights suggested by various indicators: per capita household consumption, clean water access, maize yields, and inorganic fertilizer usage. This produces the zonal sample sizes in the first column of Table 2.

A final complication is how to combine the two overlapping dimensions of stratification discussed here: (i) the rural and urban stratification and (ii) the zonal stratification. The rural and urban proportion in each zone is assigned as follows.

Let n_{rz} denote the sample size in region r of zone z, N denotes the pop-

ulation and superscript R denotes the rural sub-sample or sub-population. The sample sizes in each regional-rural/urban cell are chosen as follows:

1. The share of the sample in rural areas, Dar es Salaam, and other urban areas is fixed at 65%, 17.5% and 17.5% respectively.

2. The share of the sample in each administrative zone is fixed at the values given in Table 2.

3. The share of the sample for each region within a zone is directly proportional to the population share of the region.

$$\frac{n_{rz}}{n_z} = \frac{N_{rz}}{N_z}$$

4. The rural share of the sample in each region is equal to the rural population share in the region scaled to yield a 65% share nationally:

$$\frac{n^R_{rz}}{n_{rz}} = 0.65 \times \frac{n_{rz}}{n} \times \frac{N^R_{rz}}{N_{rz}} \div \frac{N^R}{N}$$

3.3 Stratification: HBS and other clusters

In 2007 the NBS completed a nationwide household survey of living standards, the Household Budget Survey (**?**). The HBS has provided the basis for national poverty estimates for 2007 as well as numerous other social statistics for the country. Because of its nationwide coverage and similar thematic content, the HBS is an obvious point of comparison for the NPS. From a sampling perspective, this comparison can be facilitated at multiple levels: by returning to HBS PSUs, by going one step further and tracking HBS households, or even further by tracking all individuals listed in the HBS household rosters.

In the first stage, the NPS will include all rural villages from the HBS sample plus a sample of 90 urban HBS clusters. Sampling protocols follow-

up interviews with HBS survey respondents to create a panel of households in year 1 of the NPS are discussed in section 4.1.

3.4 Other sub-samples of interest

While the NPS sample is stratified by administrative zones, it is of interest for the agricultural sector to examine crop yields and other farm-level outcomes by agro-ecological zone. Table 3 computes estimated sample sizes and standard errors for the NPS for various crops using the 2002/'03 Agricultural Census.

The standard errors shown in table 3 allow for clustering as follows. First, the ag census data is used to compute the percent of clusters in a given zone with at least one household farming the given crop. Second, the same data is used to compute the average number of households farming the crop, conditional on one household farming it. This yields the total number of clusters per crop and zone, as well as the average households per cluster. These two stages take into account both the clustering of farming activities as well as the clustering of yields for a given activity.

4 Second stage sampling: Selecting households within the cluster

The second stage of sampling involves selecting households within a given cluster. The population of households within an enumeration area or village will be defined by a household listing to be carried out in each sample cluster in the days immediately preceding the survey.

Table 3: Crop Yield (Kg/Ha): Sample Projections by Agro-Ecological Zone

		All	Alluvial	Arid	Coastal	N. High.	Plateau	Semi-Arid	S & W High.
%	HHs	100	2.8	8.2	15.8	7.4	25.8	16.2	23.9
Any cereal									
	N	1,714	42	142	283	125	456	282	420
	mean	328.2	181.4	383.6	199.8	395.6	466.3	353.9	282.3
	RSE	0.05	0.32	0.17	0.13	0.11	0.06	0.11	0.11
Maize									
	N	1,583	37	136	245	118	437	281	360
	mean	347.5	205.7	386.8	192.4	403.0	497.0	355.5	257.2
	RSE	0.05	0.31	0.17	0.13	0.13	0.06	0.11	0.11
Paddy									
	N	352	11	5	69	44	41	9	213
	mean	360.9	136.1	930.8	321.9	423.2	512.0	708.4	345.0
	RSE	0.08	0.57	0.35	0.18	0.16	0.15	0.37	0.13
Sorghum									
	N	256	10	21	84	14	30	9	28
	mean	216.3	117.4	352.4	137.5	350.2	325.8	175.5	186.1
	RSE	0.09	0.52	0.24	0.16	0.25	0.19	0.42	0.31
Cassava									
	N	330	105	5	23	115	61	9	7
	mean	203.8	239.9	171.3	247.0	97.7	351.4	184.0	220.5
	RSE	0.13	0.39	0.95	0.36	0.59	0.28	0.56	0.25
Beans									
	N	607	1	49	20	48	300	207	52
	mean	165.9	.	152.6	143.0	160.6	183.4	144.4	145.9
	RSE	0.04	.	0.16	0.24	0.13	0.05	0.08	0.16
Cotton									
	N	93	0	2	46	13	3	.	1
	mean	206.6	163.5	270.9	172.0	248.2	262.0	.	164.0
	RSE	0.10	.	0.49	0.16	0.20	.	.	1.19
Coffee									
	N	25	0	6	0	1	8	17	.
	mean	5.1	5.8	2.7	0.0	7.4	5.2	5.8	.
	RSE	0.77	.	1.22	.	.	1.84	0.66	.
Tea									
	N	3	0	.	.	.	3	2	.
	mean	278.6	1380.0	.	.	.	266.5	170.6	.
	RSE	0.94	.	.	.	.	1.06	0.66	.

Calculations are based on the 2002/03 National Sample Census of Agriculture. N refers to the number of households and RSE denotes relative standard error, or the standard error of the mean divided by the mean.

Table 4: Overview of the NPS Sample Design

	HBS Clusters		Total Sample	
	PSUs	HHs	PSUs	HHs
Mainland				
Rural	140	1,120	228	1,824
Urban	60	480	122	976
Total	200	1,600	350	2,800
Zanzibar				
Rural	..	..	30	240
Urban	..	..	30	240
Total	..	..	60	480
Grand Total	200	1,600	410	3,280

4.1 Stratification: previously interviewed versus new households

Within non-HBS clusters, a simple random sample of households will be drawn. Within HBS clusters a stratified random sample of households will be selected. One stratum will be drawn from the HBS sample to create a panel of HBS households. In either case, the NPS sample will constitute a representative cross-section of the village or EA in 2008. The representativeness of this cross-section will not depend on tracking individuals or households who have left the village or EA.

The sampling procedure within clusters will be robust to two forms of non-random attrition from the 2007 HBS sample. First a non-random group of households from the HBS sample will have relocated or simply be impossible to trace by 2008. Second, within the remaining HBS households a non-random group of individuals will have similarly disappeared. We assume that, where present, this attrition will inevitably bias population estimates based on the follow-up sample as the characteristics driving the non-random attrition will be largely unobservable.

To draw a sample which both links to the 2007 HBS and provides a representative cross-section of the 2008 population we proceed as follows:

1. The NPS will return to HBS clusters as these comprise a stratified random sample of mainland Tanzania. The NPS sample will exhaust the 140 rural villages in the 2007 HBS sample and use a sample of 90 urban EAs from the urban HBS sample.

2. Upon arriving in a village or EA, the NPS enumerators will conduct a new listing, creating an updated sampling frame of all households currently residing in the village or EA as of 2008.

3. This listing will identify households which were present in the village/EA in late 2006 at the time of the original HBS listing. For the purposes of the listing a household will be considered to have been present in 2006 if *any* member of the current household was present in the village/EA in 2006 – whether as head or as a dependent, and whether in the same or a different household.

4. Separately, the survey team will compile a list of current households containing any individual listed on an HBS roster. This HBS-specific listing or 'tracking' exercise will focus on a random sample of 12 out of the original 24 households interviewed during the HBS. For these 12 households, the current residence of all HBS household members will be recorded (if known by anyone in the village). As households may split, the list will likely contain more than 12 current households. However, only households that remain in the village or EA will be considered. Together these splintered, formerly HBS households that remain in the cluster comprise the sampling frame for the HBS follow-up interviews.

5. The 2008 population of households from the new listing will be divided into two strata: (i) those present in 2006, and (ii) newly-formed households and/or recent arrivals. For the first group, the HBS households provide a random sample. For the second group, a new random sample can be taken.

A total of eight households will be surveyed in each cluster. Of these, the proportion drawn from the HBS sample is calculated as follows:

$$\text{HBS households} = 8 \times \frac{\#\text{ of households present since 2006}}{\text{total }\#\text{ households in listing}}$$

This number of HBS households will be drawn from the list compiled in step 4 above, up to a maximum of 8. In the event that by 2008 fewer than eight households in the cluster contain members of any of the randomly chosen 12 HBS households, other households from the first stratum (resident since 2006) will be chosen at random to fill the required sample size for the stratum.

The motivation behind this sampling approach is threefold:

1. To create a panel of households linked to the 2007 HBS.
2. To avoid the expense of tracking households that have exited the village or EA.
3. To ensure that the baseline of the NPS constitutes a representative cross-section of Tanzanian households.

Under the procedures outlined here, all three of these objectives can be met. Because a new listing will be conducted to identify 'movers' and 'stayers' in the population at large, and the sample sizes will be weighted to reflect the relative sizes of these two strata, the representativeness of the baseline cross-section remains intact – even in the presence of non-random attrition from the HBS sample, and even when this attrition is driven by unobservable household characteristics.

5 Sequencing and Seasonality Issues

The NPS will be enumerated by seven mobile teams covering the entire country over a span of ten months from August to May of each year. Thus, some households will be interviewed in August, others in December, others

Table 5: Overview of NPS Mainland Sample by Region

	Region % of total pop	Rural % of regional pop	Region % of NPS sample	NPS Clusters		Of Which:			
				Total Rural	Total Urban	HBS Rural	HBS Urban	Ag Census Rural	"New" Urban
Arusha	4.1	68.4	3.7	8	5	5	2	3	3
Dar es Salaam	7.9	5.5	20.0	9	61	3	30	6	31
Dodoma	5.5	86.3	3.1	9	2	7	1	2	1
Iringa	4.8	81.8	4.6	11	5	9	2	2	3
Kagera	6.6	93.4	4.3	14	1	8	0	6	1
Kigoma	5.4	82.0	3.7	10	3	4	2	6	1
Kilimanjaro	4.5	78.0	3.7	10	3	6	2	4	1
Lindi	2.5	82.6	5.4	16	3	6	2	10	1
Manyara	3.4	83.9	2.9	8	2	5	1	3	1
Mara	4.5	80.5	2.0	5	2	5	1	0	1
Mbeya	3.4	90.3	5.4	17	2	10	1	7	1
Morogoro	5.8	71.7	4.0	9	5	7	2	2	3
Mtwara	3.7	79.4	7.1	19	6	7	3	12	3
Mwanza	9.5	79.3	4.6	11	5	9	2	2	3
Pwani	2.8	76.9	2.3	6	2	6	1	0	1
Rukwa	3.7	82.4	3.1	8	3	5	2	3	1
Ruvuma	3.6	84.8	4.9	13	4	6	2	7	2
Shinyanga	9.0	90.4	4.9	15	2	11	1	4	1
Singida	3.6	84.8	2.0	6	1	5	0	1	1
Tabora	5.7	86.5	4.0	12	2	7	1	5	1
Tanga	5.3	81.6	4.3	12	3	9	2	3	1
Total	100		100	228	122	140	60	88	62
				350		200		150	

in February, etc. Any attempt to measure production or consumption at a point in time must account for seasonality in both of these activities. The nearly year-round survey schedule of the NPS is designed to address at least two types of concerns with regard to seasonality.

First, on the production side, there is a concern for accuracy of recollection. Households interviewed just after the long rainy season harvest may have more accurate recollection of yields than those interviewed some months later. For the purpose of measuring output from the long rainy season, it would be optimal to conduct the survey for the entire country just after this harvest. Similarly, to measure output from the short rainy season, the entire survey should be timed accordingly. And to get the most accurate possible picture of planting activities, the survey would be at yet another date. The extended fieldwork of the NPS represents a compromise between all these objectives. Some households will have an immediate recollection of planting, others of fertilizer usage, others of harvest, and so on.

A second issue of seasonality relates to fluctuations in consumption. Actual expenditure over the past seven days will vary from month to month. While the issue with production is a matter of accuracy, seasonality in consumption is a question of what is being measured. Rather than producing a national poverty estimate for a single month, the NPS will seek to provide an annual estimate by averaging across months.

The key to this year-long survey strategy is that the order of interview should not be correlated with any other variables of socio-economic importance. For instance, if villages close to district capitols were interviewed early in the year, and remote villages later in the year, it would subsequently be impossible to disentangle location effects from seasonal effects. With a sufficiently large sample size, however, and by ensuring a random ordering of villages for enumeration (subject of course to certain logistical constraints) this should not pose a problem for the analysis.

? ?

6 Weights

This section presents the derivation of sampling weights. Weights are based on the probability of selection for a given cluster, household, or individual. Computation of weights for the NPS is somewhat complicated by the fact that the sample is drawn from three distinct sampling frames. First, a portion of the clusters was selected from the Population and Housing Census, with probability proportional to size (PPS) based on the population of individuals in the cluster. Second, a portion of clusters was drawn from the Household Budget Survey. Within each stratum, these HBS clusters were drawn through simple random sampling, as the original HBS sample of clusters was itself drawn through PPS from the PHC. Third, a portion of clusters was taken from the 2002 National Sample Census of Agriculture. As with the HBS sub-sample, these clusters were drawn through simple random sampling as PPS was already used to construct the NSCA sample.

The following notation will be used below. Subscripts m will denote individuals, i households, j clusters, and k strata. Capital N will indicate a population size and lowercase n a sample size. Superscripts will denote the units under consideration, such that N_j^i is the population of households in cluster j. Finally, because the HBS and NSCA samples are used as sampling frames for the NPS sampling, I will refer to these original samples as the 'population' of HBS clusters, the 'population' of NSCA households in a given stratum, etc. These frame-specific populations are denoted with superscripts in parentheses, such that $N_k^{j(HBS)}$ is the population of HBS clusters in stratum k.

6.1 Adjustment for re-weighting across strata

As discussed in detail in section 3, first stage sampling (of clusters) was stratified along several dimensions, including region and the rural/urban divide. In order to equalize standard errors across strata, those strata exhibiting

greater heterogeneity in a number of key indicators were allotted greater sampling weight. This over- or under-sampling must be accounted for in constructing cluster weights for analysis.

The adjustment factor is given by:

$$A_k = \frac{n^j}{N^j} \times \frac{N_k^j}{n_k^j} \tag{5}$$

or the proportion of the clusters in the population that were sampled, times the inverse of the proportion of the clusters in the stratum that were sampled.

6.2 Probability of selection for clusters within a stratum

Setting these adjustment factors aside for the moment and focusing within strata, there are three components to the sampling weight for a given cluster within a given stratum:

1. the probability of selection in the original sampling frame (i.e., the HBS or the NSCA),
2. the probability of selection for the NPS from among the clusters in that sampling frame, and
3. an adjustment factor based on the discrepancy between actual, current population in the cluster as estimated by the household listing and the population numbers from the PHC used in the PPS sampling.

The product of these three elements gives the final probability of selection.

6.2.1 Probability of selection into the original frame

For each of the three frames used, respectively, the probability of selection into the original sampling frame is written as:

$$P_j^{HBS} = N_j^i / N_{k(HBS)}^i \tag{6}$$

$$P_j^{NSCA} = N_j^i / N_{k(NSCA)}^i \quad (7)$$

$$P_j^{PHC} = 1/(N_k^j - N_k^{j(HBS)} - N_k^{j(NSCA)}) \quad (8)$$

Equation (6) is simply an instance of PPS sampling: it shows that the probability that a cluster was selected for the HBS was equal to the population of individuals in the cluster over the total population in the HBS stratum. The exact same explanation applies to equation (7) for the case of the NSCA. Lastly, equation (8) simply states that all remaining clusters (not sampled via HBS or NSCA) are included in the PHC sampling frame.

6.2.2 Probability of selection into the NPS sample of clusters

Since PPS was already applied to the selection of clusters for the HBS and NSCA, a simple random sample of these clusters was taken for the NPS within each stratum. Thus the probability of selection in the NPS is given by the number of clusters sampled from a given source over the total clusters in the stratum from that source:

$$P_{j(HBS)}^{NPS} = \frac{n_k^{j(HBS)}}{N_k^{j(HBS)}} \quad (9)$$

$$P_{j(NSCA)}^{NPS} = \frac{n_k^{j(NSCA)}}{N_k^{j(NSCA)}} \quad (10)$$

For the clusters taken from the PHC frame, PPS sampling was applied in the selection of clusters for the NPS. Thus the equivalent probability is:

$$P_{j(PHC)}^{NPS} = N_j^i / (N_k^j - N_k^{j(HBS)} - N_k^{j(NSCA)}) \quad (11)$$

Once again, the denominator in expression (8) highlights that sampling is done without replacement, i.e., clusters already sampled for the HBS or NSCA are removed from the PHC frame.

6.3 Probability of selection for households

After the household listing is completed, selection of households within a cluster by the field teams is done through two distinct methodologies. In both cases, all households in the cluster have an equal chance of entering the sample.

In clusters drawn from the PHC or NSCA frames, a simple random sample of 8 households is taken. In HBS clusters, by contrast, there are two complications to address. First, the sampling of households within a cluster for the HBS 2007 was stratfied by income level, based on an asset vector collected during the household listing. However, sample sizes were not re-weighted across strata; i.e., the proportion of high-, middle-, and low-income households in the sample is the same as in the population, as reflected in the household listing.

Second, as detailed in section 4.1, a portion of the sample in HBS clusters is drawn from households sampled for the HBS, creating a household panel. Nevertheless, the sample remains 'self-weighting' in that the proportion of HBS and non-HBS households was calculated to reflect the proportion of households in the listing which had been resident since the masika harvest 2006 (the time of the HBS household listing) and those which had entered the cluster since that date.

Thus, for all households in all clusters of the NPS, the probability of selection is simply

$$P_i = P_j \times \frac{n_i^j}{N_i^j} \tag{12}$$

i.e., the sample size (eight) over the total number of listed households.

6.4 Final formulae

6.4.1 Household weights

Combining the results from the previous section, we can now calculate the overall probability of selection into the NPS sample for a given household (in

a given cluster and stratum). This probability is

$$P_i = A_k \times P_j^f P_{j(f)}^{NPS} \tag{13}$$

or the product of the adjustment factor (A_k), the probability of inclusion in one of the three sampling frames (P_j^f, where f = HBS, NSCA, or PHC), and the probability that the cluster was selected for the NPS sample from within a given frame ($P_{j(f)}^{NPS}$).

6.4.2 Individual weights for household-level data

The sample weight for an individual depends on the question being asked. Since households are selected without any weighting for household size, in computing individual-level statistics based on the full sample of individuals, sampled individuals in smaller households will be over-represented. The same phenomenon occurs – but to a varying degree – in computing statistics for sub-populations, such as children of school age. Finally, an even greater tendency to over-sample individuals in small households occurs in the governance module (Section H) of the household questionnaire, where a single adult from within each household is sampled at random.

First consider an individual level statistic, computed over the entire sample of individuals, such as average height in the population. In this case, the probability of selection for an individual is identical to the selection probability of the household of which (s)he is part, and the weights used are simply the household weights.

Now, consider the slightly different case of a statistic computed at the household level, but measured in individual terms. The proportion of the population (individuals) living below the poverty line is a primary example. (Consumption data is collected at the household level, but the 'headcount' poverty line, as its name implies, counts individual heads.) In this case, the

weight given to each household should be multiplied by the household size.

$$\omega_i^m = n_i^m / P_i$$

Here the subscript denotes that this weight is to be applied to households (i) when calculating individual level statistics (superscript m).

6.4.3 Individual weights for individual-level data

6.4.4 Individual weights for the governance module

$$P_m = P_i \times \frac{n_i^j}{N_i^j}$$