

**Consultancy Report on
The World Bank Mission to Uganda
For
Developing a Sample Design for the Uganda Agriculture Census and Surveys**

**Prepared
For
The World Bank
1818 H Street, NW
Washington, DC 20433
USA**

**Ghulam Hussain Choudhry
April 28, 2008**

Table of Contents

Acknowledgement.....	3
List of Acronyms.....	4
Executive Summary.....	5
1. Introduction.....	7
1.1 Objective of the Consultancy	
1.2 Deliverables	
1.3 Activities Undertaken	
2. Agriculture Data.....	11
2.1 Need for Agriculture Data	
2.2 Current Data Sources	
2.3 Observations about the Current Situation	
3. Sample Design for the Agriculture Surveys.....	16
3.1 Dual Frame Sample Design	
3.2 Sample Size and Sample Allocation across Districts	
3.3 Sample allocation within Districts to Stages of Sampling	
3.4 Selection of EAs with PPS Systematic Sampling Procedure	
3.5 Sample of Households for the Core Module	
4. Sample Weighting and Estimation.....	33
4.1 Sampling Weights	
4.2 Survey Estimates	
4.3 Variance Estimation	
5. Conclusions and Recommendations.....	44
6. References.....	46
Annexes.....	48
Annex 1: Terms of Reference	
Annex 2: List of Officials Met and Meeting Notes	
Annex 3: Allocation of Sample (number of EAs) across Districts	
Annex 4: Probability Proportional to Size (PPS) Systematic Sampling	

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Mr. Seth N. Mayinza, Director, Division of Agriculture Statistics, UBOS, and all UBOS staff for their valuable support and collaboration during my consultancy visit to Uganda.

LIST OF ACRONYMS

AH	Agriculture Household
AM	Agriculture Module
BR	Business Register
CV	Coefficient of Variation
DFID	Department for International Development
EA	Enumeration Area
FAO	Food and Agriculture Organization
FAS	Food and Agriculture Statistics
MAAIF	Ministry of Agriculture, Animal Industry and Fisheries
MDG	Millennium Development Goals
MOS	Measure of Size
NCAL	National Census of Agriculture and Livestock
PASS	Permanent Agricultural Statistics System
PCA	Pilot Census of Agriculture
PEAP	Poverty Eradication Action Plan
PHC	Population and Housing Census
PLS&IF	Private Large Scale and Institutional Farms
PMA	Plan for Modernization of Agriculture
PPS	Probability Proportional to Size
PSFU	Private Sector Foundation Uganda
UBOS	Uganda Bureau of Statistics
UCA	Uganda Census of Agriculture
UNDP	United Nations Development Programme
UNHS	Uganda National Household Surveys
WCA	World Census of Agriculture

EXECUTIVE SUMMARY

We evaluated the sample design for the Uganda Census of Agriculture (UCA) 2008/09 surveys that the UBOS is planning to conduct. The proposed sample design consists of a dual frame design with a List Frame for the Private Large Scale and Institutional Farms and an Area Frame for the small and medium scale household-based holdings. The List Frame will be enumerated on a 100 percent basis and a sample of households will be selected from the Area Frame using a two-stage sample design with sampling of Enumeration Areas at the 1st stage and sampling of households from the selected EAs at the 2nd stage. We agree with the 100 percent enumeration of the large farms because there are very few of these but their contribution to the agricultural activity is significant. The sample planned for the Area Frame was to select 3,200 EAs, and then sample 15 households from each selected EA resulting in a total sample of 48,000 households. We determined the optimum allocation of sample to the two stages of sampling (i.e. number of EAs and number of households per EA) on the basis of cost and variance consideration. Based on the results of our analysis, we recommend a sample of 3,612 EAs and sampling of 10 households per EA resulting in a total sample of 36,120 households. This would reduce the data collection cost by 5.3 percent, and the variance of the estimates of all geographic domains (national, regional and district) would decrease by a factor of 1.056. Thus, the overall cost-variance efficiency of the recommended alternative relative to the proposed sample will be 111.4 percent. We should also note that the project budget was in deficit, and the recommended alternative turned it into a balanced budget.

The sample planned for the Core Module was the same as that for the Supplementary Module. According to the WCA 2010 recommendations the Core Module should be conducted on a 100 percent basis. If this was not feasible then the Core Module must be conducted on a “large” sample. We considered a number of options for the Core Module and because of the current budgetary constraints we are recommending that the Core Module should be conducted with the 2010 Population and Housing Census as a

piggy-back on a 100 percent basis. But a number of key items for which data would have been collected with the Core Module should be added to the questionnaire for the Supplementary Module of the UCA 2008/09.

The scheduled start date for field data collection is September 2008. We find that the proposed schedule is very tight given the number of tasks that still need to be completed before data collection operation can be started, e.g. sample selection, developing and printing enumerator and supervisor manuals, enumerator training, etc. Data Quality has been an issue during the past agriculture surveys. The main reasons for the issues related to data quality were insufficient enumerator training and lack of QA checks. Therefore, UBOS must proceed with caution so that quality is not compromised because of tight project schedule. We would emphasize that the enumerator training be strengthened and QA checks be implemented during all phases of the survey operation.

1. INTRODUCTION

1.1 Objective of the Consultancy

Uganda Bureau of Statistics (UBOS) is planning to conduct the Uganda Census of Agriculture (UCA) 2008/09. With financial support from the Department for International Development (DFID) technical assistance is being provided through the World Bank to help UBOS develop and implement an efficient sampling strategy to collect agriculture data. The terms of reference of the consultancy mission are presented in *Annex 1*. The main objectives of the consultancy mission were to:

- Develop sample design for the UCA 2008/09;
- Develop weighting and estimation methodology for the survey data; and
- Provide training to the UBOS technical staff in survey methodology.

1.2 Deliverables

The main deliverable of the consultancy mission was to develop an efficient sample design for the integrated program of UBOS Agriculture Surveys following the WCA 2010 guidelines. The sample design efficiency must be based on cost-variance optimization. Moreover, an efficient weighting and estimation methodology had to be developed that would also be simple to implement. Although, not the primary objective of this mission some training in survey methodology was to be provided to the UBOS technical staff for the purpose of capacity building. Finally, the current technical report was also one of the deliverables.

1.3 Activities Undertaken

The following activities were undertaken in connection with the consultancy mission:

1.3.1 Reviewed Reports

- Reviewed the background documents related to the UCA 2008/09.
- The documents read are included in the list of references in section 6.

1.3.2 Developed Sample Design

- Developed an optimum dual frame sample design for the UCA 2008/09 consisting of a two-stage design for the small and medium scale household-based holdings, and 100 percent enumeration of the Private Large Scale and Institutional Farms (PLS&IFs).
- Conducted research to determine sample allocation of small and medium scale household-based holdings across districts. The allocation of sample across districts is given in *Annex 3*.
- Provided mathematical formulation to determine optimum number of agricultural households to be sampled per EA on the basis of cost and variance consideration. Used the above cost-variance optimization to determine the optimum allocation of sample of households to the two stages of sampling for the UCA 2008/09.
- Provided algorithm for selecting the sample of EAs with PPS systematic sampling procedure.
- Developed and tested the SAS code for selecting the sample of EAs using the PPS sampling procedure.

1.3.3 Developed Weighting and Estimation Methodology

- Developed weighting methodology for the two-stage sample design for the small and medium scale household-based holdings.
- Provided methodology to produce survey estimates.
- Provided methodology specifications for implementing the JK2 Jackknife variance estimation.

1.3.4 Provided Training in Survey Methodology

Objective of Training Programme

Although training was not the primary objective of the current consultancy mission, it was important to provide some basic training in survey methodology for the purpose of capacity building. The format of the short training course was a Power Point presentation

during three sessions on April 03, 08 and 10 for about two hours each. The training was aimed at capacity building, and in particular upgrading the workplace skills and expertise of the technical staff engaged in the data production cycle including sample selection, data processing and analysis, and data dissemination and use.

Participants

A number of Senior Statisticians and Statisticians from three divisions of the UBOS (Division of Agriculture Statistics, Division of Socio-economic Surveys, and Division of Population and Social Statistics) and one Statistician from the Ministry of Agriculture, Animal Industry and Fisheries participated in the training.

Topics Covered

The following topics were covered during the 3 training sessions.

- Target Population
- Sample Surveys versus Censuses
- Data Quality
 - Sampling Errors and Non-sampling Errors
 - Quality Assurance Procedures
- Probability Sampling Designs
- Sampling Frames
- Basic Sampling Procedures
 - Simple Random Sampling
 - Systematic Sampling
 - Probability Proportional to Size (PPS) Sampling
 - Cluster Sampling
- Stratified Sampling Designs
- Single-stage versus Multi-stage Sampling Designs
- Stratification and Sample Allocation
- Sample Weighting
 - Base Weights
 - Non-response Adjustment
 - Post-stratification
- Error in Censuses and Sample Surveys
- Variance Estimation
 - Taylor-series Linearization
 - Replication Methods

- Standard Error of an estimate
- Coefficient of Variation (CV) of an estimate
- Confidence Interval
- Design Effect
- Definition and use of effective sample size

2. AGRICULTURE DATA

2.1 Need for Agriculture Data

Agriculture sector is the most important sector of the Ugandan economy. According to the Population and Housing Census (PHC) 2002, the agriculture sector accounted for 77 percent of the total employment for the persons aged 10 years and above. In addition, 74 percent of the households had an agricultural holding as determined from the Agriculture Module (AM) that was conducted as a piggy-back onto the PHC 2002. Therefore, information on the agriculture sector of the economy is crucial to monitoring indicators of the Millennium Development Goals (MDG). The MDGs comprise a framework of 8 goals, 18 targets and 48 indicators to be used to assess progress between 1990 and 2015, when targets are expected to be met. One of the goals is implementing the Plan for Modernization of Agriculture (PMA) in line with the Poverty Eradication Action Programme (PEAP). For more information on the MDG indicators, refer to *Indicators for Monitoring the Millennium Development Goals – Definitions, Rationale, Concepts and Sources* (UN, 2003).

2.2 Current Data Sources

The latest available agricultural data is from the Agricultural Module that was included in the 2005/06 Uganda National Household Survey (UNHS) conducted by UBOS. The Agricultural Module was included in the 2005/06 UNHS programme due to paucity of Food and Agricultural Statistics (FAS). The main objective of the 2005/06 UNHS Agricultural Module was to collect high quality and timely data on the farm economy. The survey was designed to focus on national, urban-rural and regional level data. The survey results were published in April 2007 (UBOS; 2007). The Agricultural Module of the 2005/06 UNHS was the third effort since the start of the household survey programme in 1989. The first and second were included in the Third Monitoring Survey (1995/96) and the UNHS 1999/2000 respectively. It should be noted that agriculture surveys conducted as Agricultural Modules with the national household surveys can provide coverage only for the household-based agricultural holdings.

UBOS also included an Agricultural Module in the 2002 Uganda Population and Housing Census (PHC). The data have been used to construct a sampling frame for selecting samples of Enumeration Areas (EAs) in a two-stage sample design for the household-based agricultural holdings. The frame will be used again to select the sample of EAs for the UCA 2008/09 (the currently planned survey) to provide coverage for the small and medium scale household-based agricultural holdings. A list frame will be constructed for the Private Large Scale and Institutional Farms (PLS&IFs).

Other UBOS surveys that have provided FAS include; the 2003 Pilot Census of Agriculture (PCA) whose aim was to test methodology and instrument, and the 2004 Pilot Permanent Agricultural Statistics System (PASS), which collected data on Crop Area and Production, and Livestock and Crop Utilization.

The other FAS data collected since 1990 are through the 1990/91 National Census of Agriculture and Livestock (NCAL), and the two follow-up annual sample surveys in 1991/92 and 1992/93 conducted by the Ministry of Agriculture, Animal Industry and Fisheries (MAAIF). The 1990/91 NCAL was conducted with funding from UNDP and technical support from FAO. The two follow-up surveys conducted by the MAAIF were funded by the government. The results from the 1990/91 NCAL and the two follow-up MAAIF surveys were not published because of data quality issues. Moreover, the programme could not be sustained without donor support.

2.3 Observations about the Current Situation

2.3.1 Project Plan

The key dates for the UCA 2008/09 are September 2008 for starting the field operations (listing, data collection, etc.), and January 2009 for data dissemination. This is a very tight schedule given the amount of work that still needs to be completed before the field work can be started. For example, the enumerator training manuals have not yet been completed. There is no document providing the Quality Assurance plan during field operations. District Officials are currently updating the BR 2006/07 listings of the large agricultural farms but the methodology for the task has not been documented. Some

consultants have even suggested delaying the field work by six months but the bureau is faced with two constraints.

1. The portion of the budget allocated to 2008/09 fiscal year must be spent during the fiscal year.
2. The second constraint is perhaps more important. The bureau has to devote time and resources to the upcoming Population and Housing Census in 2010, and cannot afford to postpone the UCA 2008/09 by another six months.

In view of the above constraints, we would suggest that UBOS proceed with extreme caution by implementing proper quality checks in spite of the tight schedule. Otherwise, the quality will be at risk if quality checks are not implemented because of lack of time.

2.3.2 Technical Documents

The agricultural questionnaires have been developed with input from the Ministry of Agriculture, Animal Industry and Fisheries, but these are not quite ready for printing. As noted by other consultants as well, the questionnaires developed so far do not follow the structure recommended by the WCA 2010, i.e. Modular and Integrated Approach.

After careful review of the questionnaires, we also found that the information about *partnership agricultural holdings* needed for sample weighting is not being collected. The sampling unit is the householder and we must determine the number of unique householders in the partnership and not the number of persons. For example, if the partnership is among persons within the same householder then there is no implication for weighting. It should be emphasized that if partnership adjustment factor is not applied during weighting the survey estimates will be subject to upward bias. Moreover, the final disposition codes for the sampled households are not being collected either. The disposition codes will be needed to determine the response status categories, which are:

- 1 = respondent,
- 2 = non-respondent, and
- 3 = out-of-scope.

The non-response categories are used to compute non-response adjustment factors during weighting. Therefore, these data items must be included on the Listing Module (UCA Form 1).

Recommendation 1: The information about *partnership agricultural holdings* needed for sample weighting must be collected. Moreover, the questionnaire disposition codes (Result Codes) must be collected as these are also needed for weighting.

The other technical document related to the UCA 2008/09 that is almost ready is the Project Document prepared by UBOS dated September 2007 (UBOS; 2007). The document specifies the objectives, main methodology, work plan and detailed budget for the 3 financial years starting from the 2007/08 financial year. **The document is currently being revised based on the recommendations made during the current consultancy mission.**

The Technical Report by Enock Ching'anda (2008) provides specific recommendations for cost-effective methods of collecting crop production data in Uganda, covering both area measurement and yield. Implementation of these recommendations will be a step forward in enhancing the agricultural data quality.

Related manuals for the field work (Enumerator's Manual, Supervisor's Manual, QA Officer's Manual, etc.) have not yet been developed. UBOS plans to update the existing Enumerator's Manual and Supervisor's Manual from the Pilot Census of Agriculture 2003 (UBOS; 2004a). Since there was no QA Officer's Manual for the PCA 2003, a QA plan and the corresponding QA Officer's Manual need to be developed for the field operations.

2.3.3 Data Quality

The software *STATA* is used to compute standard errors of the estimates using linearization method, which accounts for the complex sample design. The CVs of the estimates are then computed for various publications. The non-sampling errors have been

mentioned in some of the publications but the quality assurance (QA) procedures to control and minimize these errors have not been documented. In general, data quality is an issue that needs attention in order to improve the agricultural data quality. For example, the number of agricultural households at the district level produced from the 2002 PHC Agriculture Module is larger than the number of households from the 2002 PHC for 3 out of 56 districts (UBOS; 2004b – Table 3.1 on page 25). This is an indication of lack of quality checks during processing.

One of the reasons for data quality issues has to do with inadequate training. In order to enhance agricultural data quality, it is proposed that training in agricultural statistics be provided to all technical staff in the data production cycle including collection, processing and analysis, and dissemination and use. The training would aim at upgrading the expertise and skills of the technical staff to enable them to develop and implement QA procedures. In addition, the field enumerator training must be very intensive so that they can apply the concepts and definitions properly.

Recommendation 2: We recommend developing and implementing QA procedures following the guidelines given in the Statistics Canada¹ (2003).

2.3.4 Further Technical Assistance

UBOS can benefit from further technical assistance during the following phases of the UCA 2008/09.

- Field data collection operation – Two weeks.
- Data processing operation – Two weeks.

¹ The publication Statistics Canada Quality Guidelines is available at the website: <http://www.statcan.ca/english/freepub/12-539-XIE/steps/coverage.htm>

3. SAMPLE DESIGN FOR THE AGRICULTURE SURVEYS

The 2002 Population and Housing Census (PHC) conducted in September 2002 included an Agricultural Module (AM) whose main purpose was to provide appropriate sampling frame for a detailed Census of Agriculture and Livestock, and other agricultural surveys. The sampling frame thus constructed would provide coverage for the household-based agricultural activity only. In order to provide complete coverage of the agricultural activity, another sampling frame will be constructed to collect information for the Private Large Scale and Institutional Farms (PLS&IFs) resulting in a **Dual Frame sample design**.

3.1 List Frame Sample

The Large Scale Private and Institutional Farms (PLS&IFs) contribute significantly to the Ugandan economy but they are small in number². Therefore, no sampling will be done and these will be enumerated on a 100 percent basis.

3.1.1 Construction of List Frame

UBOS in collaboration with the Private Sector Foundation Uganda (PSFU) undertook the update of the Business Register (BR) that was created during 2001/02. The setup of the 2006/07 BR is such that the sectors are considered as they appear in the International Standard Industrial Classification (ISIC). The agriculture sector that is of interest to the UCA 2008/09 surveys covers crop growing including fruits and vegetables, and livestock including sheep, goats, etc. A distribution of the agricultural businesses by type shows that 56 percent of the businesses were engaged in Livestock Agriculture followed by 24 percent in Mixed Farming (i.e. both Livestock and Crop), and the rest were only in Crop growing.

The District Production Coordinators are currently updating the lists of the PLS&IFs using criteria developed during the 2003 Pilot Census of Agriculture (see

² Based on the 2006/07 BR their number is 412 Farms.

UBOS; 2004a). The field updates consist of verifying the agricultural businesses on the BR 2006/07, and adding new businesses that qualify under the above criteria based on the local knowledge.

Recommendation 3: It is recommended that the respective supervisors in their districts check a sample of agricultural businesses by field visit after the listings have been updated.

3.2 Area Frame Sample Design

A stratified two-stage sample design will be used for the small and medium scale household-based agricultural holdings (Area Frame design). At the first stage Enumeration Areas (EAs) will be selected with Probability Proportional to Size (PPS) systematic sampling, and at the second stage households which are the ultimate sampling units will be selected with systematic sampling. A total of 3,833,485 out of the 5,186,558 households enumerated during the 2002 PHC reported that one or more of their members were engaged in an agricultural activity as of September 2002. These households will be referred to as “households with agricultural activity” or “agricultural households”. The advantage of the data from the AM is that it was collected on a universal basis (*complete enumeration of households*). The major drawbacks of the AM data are:

- The AM questionnaire was very brief compared to those designed for conventional agricultural surveys/censuses. As a result, very important questions on agriculture had to be left out in order to keep the AM questionnaire short.
- There were numerous data quality problems because of insufficient enumerator training, and lack of rigorous Quality Assurance (QA) checks during data collection.
- The questions on the agricultural activities did not have adequate filters to differentiate between activities within the EA where the household was located or outside the EA or even outside the district.

Since there is no other data source to construct a sampling frame for the household-based agricultural holdings we will use the list of EAs as the sampling frame to select the sample of EAs with PPS systematic sampling using the number of “Agricultural Households” collected during the 2002 PHC as the measure of size (MOS). It should be noted that the 2002 PHC data will be used only as MOS for the PPS sampling of EAs. But the sampled EAs will be listed in the field and a number of filter questions will be administered to determine eligibility (*i.e., only the Households with Agricultural Activity and not covered by the List Frame will be eligible*). It is important to note that the agricultural households that are on the list frame (PLS&IFs) will be ineligible for the area frame sample. Therefore, sufficient information will be collected on the Listing Module (UCA Form 1) to determine whether the household was already included in the List Frame or not. It should be noted that any household with agricultural activity that is not on the List Frame will be eligible for the area frame sample even if it satisfied the criteria of Private Large Scale holding. In other words, the List Frame cannot be updated based on the information collected from the EAs sampled from the Area Frame.

The Core Module questionnaire will be completed for all Agricultural Households in the sampled EAs (see section 3.4). The information collected with the Core Module questionnaire will be used at the design stage (e.g. to determine eligibility for the Supplementary Module, size stratification, etc.) to select the sample of agricultural households for the Supplementary Module. Moreover, the auxiliary information collected with the Core Module questionnaire from a much larger sample of households can be used at the estimation stage to improve the precision of the survey estimates produced from the Supplementary Module.

3.3 Sample Size and Sample Allocation for the Area Frame

The survey estimates from the 2008/09 Uganda Census of Agriculture (UCA) will be produced at the national, regional and district levels. The country is divided into 80 districts, and there is large variation in the size of the districts where size is the number of

Agricultural Households. There are four regions, which are statistical regions and not the administrative regions and these are defined as groupings of the districts.

In order to produce reliable survey estimates for each of the 80 districts the sample size for the household-based holdings will be quite large. UBOS plans to select 48,000 agricultural households from the small and medium scale household-based holdings (Area Frame). The total sample of 48,000 agricultural households will be allocated in two steps. First, the total sample will be allocated to the 80 districts. Then, the sample within each district will be allocated to the two sampling stages, i.e. number of EAs to be selected, and number of households to be selected from each sampled EA.

3.3.1 Allocation of the Area Frame Sample across Districts

The distribution of the number of Agricultural Households obtained from the 2002 PHC will be the basis for allocating the Area Frame sample across districts. We consider three approaches to allocate the total sample across 80 districts: (1) equal allocation, (2) proportional allocation, and (3) compromise allocation.

Approach I – Equal Allocation: Under this approach, the sample is allocated equally to each of the districts. The equal allocation approach would achieve roughly the same reliability for the district level estimates. Because of the large variation in the size of the districts, choosing this approach would result in large variation in the selection probabilities of households across districts. As a result, the variation between the sampling weights would be very large and would result in large variances for the regional and national level estimates. We therefore would not choose this allocation because it would have adverse impact on the precision of the national and regional level estimates.

Approach II – Proportional Allocation: Under the proportional allocation, the larger districts would receive the larger share of the sample. Although, the proportional allocation would be the most efficient allocation for the national level estimates, the estimates for the smaller districts would not be very reliable. Therefore, this allocation cannot be considered either.

Approach III – Compromise Allocation: As the name implies, the compromise allocation is aimed at striking a balance between producing reliable district level estimates (*Approach I*) and reliable national level estimates (*Approach II*). A number of procedures are available to achieve this compromise. The simplest and most commonly used allocation is the so-called “square root” allocation. A more general compromise allocation is the “power allocation” discussed by Bankier (1988) under which the sample is allocated proportional to x^λ , where x is the measure of size (MOS) and the parameter λ can take values between zero and 1. The value $\lambda = \frac{1}{2}$ corresponds to the “square root allocation.” The two extreme values of λ give the “equal allocation” and the “proportional allocation.” In other words, $\lambda = 0$ corresponds to *Approach I*, which is “equal allocation” and $\lambda = 1$ corresponds to *Approach II*, which is “proportional allocation.” Kish (1988) has also considered a number of compromise allocations including the “square root” allocation.

Because we are interested in both national level estimates and the estimates for each of the districts, we computed the design effect for the national level estimates for different power allocations. It should be noted that we will not control the sample size by region because the regions are aggregations of the districts, and hence are quite large.

We provide in Table 3-1 the design effects for the national level estimates due to variation in design weights, and the corresponding sample size for the smallest and largest districts (sample size range) for different “power allocations” with $\lambda = 0.0, 0.1, 0.2, 0.3, \dots, 0.9, 1.0$. The district level sample sizes are computed based on the total national sample size of 48,000 households. It should be noted that the current analysis is only for the purpose of determining the sample allocation across districts. As discussed later in sub-section 3.3.2, the total national sample size will also be revised.

The design effect due to variation in the design weights is defined as $(1 + CV^2)$, where CV is the coefficient of variation of the design weights. As pointed above, $\lambda = 0$ is the equal allocation, and $\lambda = 1$ is the proportional allocation.

Table 3-1: Design Effect of National estimates due to variation in the weights, and Sample Size Range for Districts for Different Power Allocations

Power	Design Effect	District Sample Size (Households)	
		Minimum	Maximum
0.0	1.30	600	600
0.1	1.23	470	671
0.2	1.18	366	747
0.3	1.13	284	828
0.4	1.10	220	915
0.5	1.07	169	1,008
0.6	1.04	130	1,106
0.7	1.02	100	1,210
0.8	1.01	76	1,320
0.9	1.00	58	1,436
1.0	1.00	44	1,558

We notice from Table 3-1 that the design effect for the national level estimates due to variation in the design weights is 1.30 for equal allocation, which is very high. The design effect for the “square root allocation” ($\lambda = 0.5$) is 1.07, which is not too high but the sample sizes for the smaller districts will be too low. For example, the sample size for KALANGALA, which is the smallest district, will be only 169 households. If we use power allocation with $\lambda = 0.4$ the design effect for the national estimates would become 1.10 and the sample size for KALANGALA would increase from 169 to 220 households, and the sample for all other districts would be greater than 300 households. Therefore,

this power allocation seems a reasonable compromise between producing reliable district level estimates without having significant adverse impact on the precision of the national estimates.

As discussed later in sub-section 3.3.2, the total sample size recommended for the UCA 2008/09 will be 36,120 households instead of the currently planned sample of 48,000 households. Therefore, the following recommendation is for the purpose of determining the distribution of the total sample across districts, and not for determining the absolute sample sizes for the districts.

Recommendation 4: We recommend power allocation with $\lambda = 0.4$ for allocating the Area Frame sample across districts. Since the sample for KALANGALA will be somewhat on the low side it should be increased to the same sample size as for KAMULI, the 2nd smallest district.

3.3.2 Allocation of Sample within Districts to Stages of Sampling

After allocating the sample across districts, we will determine optimum allocation of the district level sample across the two stages of sampling. For example, a sample of 450 households for a district can be selected by selecting 45 EAs and then selecting 10 households from each sampled EA. Alternatively, we can also select 450 households by selecting 30 EAs and then selecting 15 households from each sampled EA. The choice of number of households to be sampled per EA determines the number of EAs that will be sampled, and it has cost and variance implications. As the number of sampled households per EA increases the number of EAs to be sampled would decrease and the resulting data collection cost would decrease. But the variance would increase because of increased clustering of the sample. We can obtain optimum EA sample (number of households to be sampled per EA) under a linear cost model as follows.

Suppose we want to sample M households from the district by selecting m households per EA. Then, the number of EAs to be sampled from the district, say n will be given by M/m with some rounding. We will use the linear cost

model $C = n \times c_1 + nm \times c_2$, where c_1 is the unit EA cost (i.e., travelling to the EA, listing the EA and screening, data capturing the listing, sampling, etc.) and c_2 is the unit cost of enumerating one agricultural holding (strictly speaking, it is the average cost of enumerating all holdings linked to one agricultural household because household is the sampling unit). The design effect due to clustering is given by $\{1 + (m-1) \times \rho\}$, where ρ is the intra-cluster correlation (or rate of homogeneity). The parameter ρ is given by the ratio:

$$\frac{SSB}{(SSB + SSW)}, \quad (3-1)$$

where SSB and SSW are respectively the between and within EA sum of squares. The value of ρ can be estimated from the past survey or census data. Since ρ will be different for different variables, an average (or median) value can be taken for few key variables. The optimum value of number of households to be sampled per EA under the above cost model is given by:

$$m_0 = \sqrt{\left(\frac{c_1}{c_2}\right) \times \left(\frac{1-\rho}{\rho}\right)} \quad (3-2)$$

After substituting the expression for ρ from equation (3-1) into equation (3-2) and simplifying, we can write the expression for m_0 as:

$$m_0 = \sqrt{\left(\frac{c_1}{c_2}\right) \times \left(\frac{SSW}{SSB}\right)}. \quad (3-3)$$

We note from equation (3-3) that the number of sampled households per EA would increase as the average cost of an EA (including travel, listing, screening, data capture, etc.) relative to the average cost of enumerating an agricultural holding for the Supplementary Module increases. On the other hand, as the variability within the EA relative to the variability between EAs decreases (i.e. households within EAs are more

similar than across EAs with respect to the variable of interest) the number of sampled households per EA would decrease. In other words, there will be no benefit from selecting too many households from the EA if they are very similar to each other.

We used the data from the AM that was piggy-backed onto the 2002 PHC to compute the values of ρ for three variables: number of Pure Plots, number of Mixed Plots, and number of Total Plots, which is the sum of the number of Pure Plots and number of Mixed Plots. We computed the values of ρ at the District level and then these were averaged. The average value of ρ was 0.27 for the variable Pure Plots, it was 0.29 for the Mixed Plots and it was 0.31 for the Total Plots. We would have also liked to compute ρ for the variable Land Area under Cultivation but the variable was not included on the AM during the PHC 2002. The variable “Land Area of Holding” by Type of Use is included as part of the Core Module for the UCA 2008/09. If the Core Module gets postponed until 2010, the variable must be included as part of the next Core Module because the data will be useful for designing the future Agriculture Surveys.

Recommendation 5: The variable “Land Area of Holding” by Type of Use must be included in the next Core Model. The variable will be useful for designing future Agriculture Surveys.

Although, the value of ρ will vary from one variable to another, the above analysis gives us an idea about the order of magnitude of ρ . We provide in Table 3-2 the optimum values of number of households to be sampled per EA for different values of the cost ratio, and the rate of homogeneity equal to 0.29 as determined above (or equivalently the ratio SSW/SSB equal to 2.45).

Table 3-2: Optimum Number of Households to be sampled per EA

Cost Ratio	Optimum Number of Households to be sampled per EA
15	6
20	7
25	8
30	9
40	10
50	11
60	12

The above analysis shows that the optimum number of households to be sampled per EA would be at most 10 households per EA even if the cost ratio was as high as 40.

Recommendation 6: UBOS is currently planning to select 15 householders per EA. We recommend that the number of households to be sampled per EA should be reduced from 15 to 10 householders.

Cost-Variance Implication

For the same given field cost we can afford to select a larger sample of EAs when selecting 10 households per EA as compared with the current plan of selecting 3,200 EAs and sampling 15 households per EA. On the variance side, the design effect due to clustering will be 5.06 when sampling 15 households per EA, and it will be 3.61 when sampling 10 households per EA. Therefore, the choice among alternative sample designs (including the total sample size) must be based on the cost-variance efficiency of the various alternatives.

First, we observe that the *effective sample size* under the current plan would be 9,486 households (48,000 divided by the corresponding design effect of 5.06). The

effective sample size under the recommended design must be at least the same as under the current plan so that we would achieve the same (or smaller) variance as compared with the current plan. In order to achieve the same effective sample size with 10 households per EA as the current plan we would need to sample about 3,425 EAs (9,486 multiplied by the corresponding design effect 3.61, and divided by 10). But, we will consider a more conservative alternative where we will select 3,600 EAs and sample 10 households per EA so that the total sample will be 36,000 households. Thus, there will be a 25 percent reduction in the sample size (i.e., the sample size will reduce from 48,000 households to 36,000 households).

Once the total sample size and the distribution of the sample across Districts are determined we can obtain the number of EAs to be sampled from each district. Let us say that the total sample is now 36,000 households and it is distributed across the 80 districts according to the power allocation ($\lambda=0.4$). The sample size for KALANGALA, the smallest district will be only 165 households, which is somewhat low and should be increased to the same size as the next larger district, which is KAMULI and has a sample of 227 households. Next, we determine the number of EAs to be selected from each district under the assumption that 10 households will be sampled per EA. The total sample size in terms of number of EAs actually becomes 3,612 EAs instead of the targeted 3,600 EAs for two reasons:

1. The sample for KALANGALA was increased by 6 EAs, and
2. The total sample also increased by another 6 EAs because of rounding.

Thus, the total sample size under the recommended alternative will be 36,120 households.

Cost-Variance Efficiency

We now determine the cost-variance efficiency (see Choudhry, Lee and Drew; 1985) of the recommended alternative relative to the currently planned sample. It can be shown that the variance efficiency of the recommended sample of 36,120 households selected by sampling 10 households per EA is 105.5 percent as compared with the sample

of 48,000 households selected by sampling 15 households per EA. On the cost side, the total data collection cost was estimated to be USG 16.639B for the planned sample of 48,000 households (sample of 3,200 EAs and 15 households per EA). We also obtained the total data collection cost of the recommended alternative (sample of 3,612 EAs and 10 households per EA), which turned out to be USH 15.753B. Thus, the cost efficiency of the recommended alternative relative to the current plan is 105.6 percent (16.639 divided by 15.753). Thus, the overall cost-variance efficiency defined as the product of the variance efficiency and the cost efficiency will be 111.4 percent. Incidentally, the project budget was in deficit and the budget got balanced because of the recommended change in the design.

Recommendation 7: Based on the results of the above analysis we recommend that a sample of 3,612 EAs should be selected for the 2008/2009 UCA, with 10 households per EA resulting in a total sample of 36,120 households. The sample of 3,612 EAs will be allocated across districts using power allocation with $\lambda = 0.4$.

The allocation of sample of 3,612 EAs across districts is given in *Annex 3*.

3.4 Selection of EAs with PPS Systematic Sampling Procedure

The sample of required number of EAs will be selected from each district with probabilities proportional to size (PPS), using the systematic sampling algorithm described in Hansen, Hurwitz, and Madow (1953). The measure of size (MOS) to be used for sample selection will be the number of Agricultural Households determined from the 2002 PHC. Probability proportional to size (PPS) sampling is an efficient procedure that is used widely in multi-stage (in this case, two-stage) sampling designs. A regular feature of such designs is that the clusters sampled at various stages of sampling are markedly unequal in size (that is, the number of elements they contain). We provide in *Annex 4* the algorithm to select the sample of EAs from each district. **It is important that the EAs be sorted by County and Sub-county within the Districts, and then by MOS by alternating between “ascending” and “descending” orders from one Sub-county to the next.**

As described in *Annex 4* the PPS sampling procedure is implemented by normalizing the measures of size so that these add up to the sample size in terms of number of EAs to be sampled from the District. The advantages of using the “Normalized Measure of Size” are:

- The random start will always be a random number between zero and 1.
- The skip interval will always be equal to 1.
- The normalized size measure will be the selection probability of the EA.

In the context of PPS sampling, the normalized size measure (i.e. selection probability) must always be less than 1. It may happen that an EA is so large that the corresponding “normalized size measure” becomes greater than 1. It is more likely to happen in the “smaller” districts than in the “larger” ones. The following two options can be considered for the “very large” EAs.

Option1: Divide the very large EA into a number of pseudo EAs by a “conceptual split” where each pseudo EA would be considered to be of the same size. The number of “conceptual splits” will be equal to 2 if $1 \leq \text{normalized-size} < 2$, it will be equal to 3 if $2 \leq \text{normalized-size} < 3$, and so on. The EA will still be one “physical” EA and a 2nd stage sample of agricultural households will be selected for each sampled pseudo EA. A “weight adjustment” will be applied to account for the “conceptual split” because the original EA would now represent two or more pseudo EAs.

Option2: Select the large EAs with Normalized-Size ≥ 1 with certainty (selection probability equal to 1). Adjust the sample size for the district, and re-compute the “Normalized Sizes” and check that there are no new EAs with Normalized-Size ≥ 1 . The weighting and estimation including variance estimation for the “certainty” EAs would be implemented to reflect the single-stage design for the large EAs.

The option 1 will be preferable because the same processing system can be used for all EAs.

Very Small EAs

It is also possible that there will be some very small EAs. It will not be very cost effective to sample a very small EA. These EAs should be collapsed with neighboring EAs before sampling. It would be preferable to do the collapsing of the small EA with a contiguous EA. If a collapsed EA gets sampled, it will be treated as if it was a single EA.

Recommendation 8: We recommend option 1 for “large” EAs because the same weighting and estimation procedures can be used for the “large” EAs as for the other regular EAs. The only additional step would be to apply “weight adjustment” to account for the “conceptual split” of the large EA. We also recommend that very small EAs (say, EAs with less than 20 agricultural households) be collapsed with another EA.

An Example to Illustrate PPS Sampling Procedure

We use the example given in Table 3-2 to illustrate the PPS Systematic Sampling procedure with 18 EAs in the stratum from which 4 were to be sampled. In order to keep the number of EAs in the example small we did not use counties and sub-counties. Moreover, the EAs were sorted in the ascending size order, and were assigned sequential numbers from EA01 to EA18 (Column 1) after the sort. The measures of size (MOS) of the EAs are given in Column 2 and the individual selection probabilities (normalized size measures) are given in Column 3. The cumulative “normalized size measures” are given in column 4.

In order to select the sample of EAs, we will generate a uniform random number between 0 and 1. Let us say the random number was 0.2192. Then, the first sampled EA will be the one with cumulative “normalized size measure” greater than or equal to 0.2192 such that the previous cumulative “normalized size measure” was less than 0.2192. Therefore, the EA number “EA02” will be selected because its cumulative “normalized size measure” is greater than 0.2192. The other 3 EAs that will be selected will be the ones with the cumulative “normalized size measures” greater than or equal to

the selection numbers 1.2192, 2.2192 and 3.2192 such that their previous cumulative “normalized size measures” were less than these selection numbers respectively. Thus, the EAs with sequence numbers “EA02”, “EA07”, “EA12” and “EA16” will be selected. The sampled EAs have been flagged with a “*” in column 5.

Table 3-3: Illustration of EA Sampling with PPS Systematic Method ($N = 18$; $n = 4$)

EA NUMBER (1)	EA MOS (2)	PROBABILITY (3)	CUMMULATIVE (4)	SELECT (5)
EA01	116	0.1561	0.1561	
EA02	122	0.1641	0.3202	*
EA03	128	0.1722	0.4924	
EA04	132	0.1776	0.6700	
EA05	134	0.1803	0.8503	
EA06	142	0.1911	1.0414	
EA07	151	0.2032	1.2445	*
EA08	163	0.2193	1.4638	
EA09	168	0.2260	1.6899	
EA10	174	0.2341	1.9240	
EA11	178	0.2395	2.1635	
EA12	182	0.2449	2.4083	*
EA13	186	0.2503	2.6586	
EA14	190	0.2556	2.9142	
EA15	192	0.2583	3.1726	
EA16	198	0.2664	3.4390	*
EA17	201	0.2704	3.7094	
EA18	216	0.2906	4.0000	

3.5 Sample of Households for the Core Module

The new FAO recommendations included in the WCA 2010 insist that the countries use a modular approach to meet the need for a wide range of data from the agricultural census. Ideally, the Core Module would be conducted on a complete enumeration basis to provide a limited range of key structural items of importance for national-policy making. The Core Module would then form the basis to construct the sampling frames for all other Agriculture Surveys. There is neither the time nor the resources to undertake such large scale data collection operation. As suggested in WCA 2010, the Core Module can also be conducted on a “large” sample basis. It should be noted that the key point is that the sample size (in terms of number of households) for the Core Module should be large so that it can be used as a basis for constructing the sampling frames for all other Agriculture Surveys. We considered three options for the sample of households to conduct the Core Module:

Option 1: Conduct the Core Module for all households in the sampled EAs. The Core Module will be conducted simultaneously with the field EA listing operation. Thus, the Core Module will be conducted for all agricultural households in the 3,612 sampled EAs.

Option 2: Do not conduct the Core Module as part of the UCA 2008/09, but include some key items of the Core Module with the Supplementary Module. This approach is not much different from the current plan because the sample for the planned Core Module was the same as the sample for the Supplementary Module. The Core Module will be conducted as a *piggy-back* onto the PHC 2010 on a 100 percent basis similar to the approach used during the PHC 2002 but unlike the AM during the PHC 2002 all key items recommended by WCA 2010 will be included in the Core Module.

Option 3: Select a large sample of households from the sampled EAs (say, 50 percent of the households) to conduct the Core Module. The sample for the Supplementary Module will be a sub-sample of the Core Module sample.

We cannot recommend **option 3** because there would be operational problems with this option. For example, sampling in the field during the field listing operation may not be feasible because it would require extensive enumerator training in sampling. Alternatively, the sample of households could be selected at the UBOS head office after the field listing operation but there may not be any (or much) cost savings as compared with that of option 1 because the enumerators would have to make a second field visit to conduct the Core Module. In either case, there will be adverse impact on the data quality (higher CVs of the estimates) due to reduced sample size for the Core Module.

Therefore, UBOS would have to make a choice between **options 1 and 2** based on the budget situation. The rationale for **option 2** is that Core Module should be only done when there is a budget to conduct the Core Module properly.

Recommendation 9: Because of the current budget situation we recommend that the Core Module should be postponed until 2010. The Core Module should be conducted in 2010 on a 100 percent basis by piggy-backing onto the PHC 2010.

4. SAMPLE WEIGHTING AND ESTIMATION

After the data collection and editing phases of the 2008/2009 National Agriculture Survey, the sampling weights for the data collected from the sampled agricultural holdings will be constructed so that the responses could be properly expanded to represent the entire population of agricultural holdings.

4.1 Sampling Weights

The weights will be the result of calculations involving several factors, including original selection probabilities, adjustment for non-response and benchmarking to the “Core Module” which would have a much larger sample than the Supplementary Module sample if it was conducted. The weights will be produced for the household-based agricultural holdings only. The weights for the Private Large Scale and Institutional Farms will be equal to 1 as these will be enumerated on a 100 percent basis.

4.1.1 Calculation of Base Weights

The base weight (or design weight) for each agriculture holding will be equal to the reciprocal of its probability of selection. The probability of selection of a holding is the product of the probability of selecting the EA and the conditional probability of selecting the holding given that the EA had been selected. The EAs are selected with PPS systematic sampling procedure. The formula for the EA selection probability is given in Step 3 of *Annex 4*. The conditional selection probability of the agricultural holding (which is linked to the households in the sampled EA) will depend on the selection probability of the household or households that the agriculture holding is linked to. It should be noted that the sampling unit is Agricultural Household, whereas the unit of observation (or data collection) is the agricultural holding. Therefore, probability of selecting a holding would depend on the selection probability of the household or households associated with the holding. The householders may be classified into the following three householder sectors:

1. Single-holding householder,
2. Multiple-holding householder,
3. Partnership of two or more householders.

The probability of selecting a holding associated with the 1st or 2nd type of householder sectors is straight forward, but a holding associated with multiple householders (3rd type of sector) has multiple roots of selection. Therefore, computation of exact selection probability becomes very complicated. A simple alternate is to prorate the reported holding level data by constructing pseudo-holdings corresponding to the multiple householders in the partnership. Otherwise, the estimates would be subject to upward bias. Equivalently, the weight of the agricultural holding can be adjusted by multiplying with a factor equal to the reciprocal of the number of householders the agricultural holding is linked to.

4.1.2 Non-response Adjustment

Properly weighted estimates using the base weights (as given above) would be unbiased if each sampled eligible agricultural holding would agree to participate in the survey. However, non-response is always present in any survey operation, even when participation is not voluntary. Thus, weight adjustment would be necessary to account for the non-respondent agricultural holdings. The non-response adjustment will be applied within each EA. Under the assumption of random non-response the weight adjustment will be the ratio of number of sampled eligible holdings and the number that actually responded. If n eligible holdings were sampled from an EA and only r were enumerated then the non-response adjustment factor will be equal to n/r . It should be noted that we have assumed that there will be no EA with very low response rate. The response rate should be at least 60 percent or more for every EA. Otherwise, the EA should be collapsed with another EA to apply the non-response adjustment. This is a trade-off between accepting higher variance because of very large non-response adjustment factors on one hand, and reducing the variance by collapsing EAs but taking the risk of higher non-response bias on the other hand.

In order to improve the reliability of the survey estimates we would also apply a post-stratification adjustment using information from the Core Module. We recall that the sample for the Core Module (if conducted) will be much larger than that for the Supplementary Module. There is a possibility that the Core Module may not be conducted at this time, in which case the non-response adjusted weights will be used for estimation.

4.1.3 Post-stratification Adjustment

Post-stratification is a popular estimation procedure in which the base weights after non-response adjustment are further adjusted so that the estimated totals based on the adjusted weights are equal to known population totals (or more precise estimates of the population totals) for certain subgroups of the population. Since sample size for the Core Module will be much larger than that for the Supplementary Module we can use estimates based on the Core Module as control totals for post-stratification.

We will define the post-strata to be the cross classification of counties (or groupings of counties) and size categories (i.e. size of the agricultural holdings; small, medium, and large). Let Z be the auxiliary variable that is collected for the Core Module and it is highly correlated with the study variable Y . The auxiliary variable Z can be a quantitative variable (e.g., land area under cultivation) or an indicator variable, i.e. yes/no with values 1 and 0 (e.g., presence or absence of irrigation system). We will denote by \tilde{Z}_g the estimated total of the variable Z for the post-stratum g ($g = 1, 2, 3, \dots, G$) that is based on the Core Module sample. Similarly, the estimated total that is based on the Supplementary Module sample will be denoted by \hat{Z}_g . The post-stratification adjustment is then defined as the ratio \tilde{Z}_g / \hat{Z}_g . If the sample size for a post-stratification cell from the Supplementary Module is very small the post-stratum should be collapsed with another post-stratum.

4.1.4 Calculation of the Final Weights

The post-stratified weights of the respondent agriculture holdings will be calculated as the product of the non-response adjusted base weights and the corresponding post-stratification adjustment. The post-stratified weights will be the final weights of the respondent agricultural holdings that will be used to construct survey estimates for various estimation domains.

Let w_i be the non-response adjusted base weight of the agricultural holding i , then the post-stratified weight w_i^* of the agricultural holding will be computed as:

$$w_i^* = \left(\frac{\tilde{Z}_g}{\hat{Z}_g} \right) \times w_i, \quad i \in g. \quad (4-1)$$

The superscript (*) is used to denote that it is a post-stratified weight. Because an agricultural holding i can belong to one and only one of the post-strata, the post-stratified weights are uniquely defined. The advantage of post-stratified weighting is that the reliability of the survey estimates is improved when there is high correlation between the auxiliary variable used for post-stratification and the study variable. Moreover, most of the bias due to under-coverage is corrected.

4.2 Survey Estimates

All survey estimates will be obtained as domain estimates by using an indicator variable ${}_d\delta_i$, where the post-script d denotes the “estimation domain” and the sub-script i denotes the respondent agricultural holding. The estimation domain can be a geographic domain (e.g., a district) or it can be a characteristic domain (e.g., female agricultural holders). The estimation domain can also be the intersection of two or more geographic and/or characteristics domains. For example, all female agricultural holders in a particular district who harvest wheat grain using an irrigation system. The indicator variable ${}_d\delta_i$ is defined as:

$${}_d\delta_i = \begin{cases} 1, & i \in d \\ 0, & \text{Otherwise} \end{cases} \quad (4-2)$$

Then the estimated total of the study variable Y for the domain of interest d can be expressed as:

$${}_d\hat{Y} = \sum_{i \in s} w_i^* \times {}_d\delta_i \times y_i, \quad (4-3)$$

where y_i is the reported value of the study variable Y for the respondent agricultural holding i , and w_i^* is the corresponding survey weight. The summation symbol $\sum_{i \in s}$ denotes the summation over all selected agricultural holdings that provided useable data. The indicator variable ${}_d\delta_i$ defined in equation (4-2) would include the contribution only from those respondent agricultural holdings that belong to the estimation domain. For example, if we were interested in the estimate of total wheat production in a particular district then the value of ${}_d\delta_i$ will be equal to 1 for the respondent agricultural holdings that are in the given district and have wheat production, and it will be 0 for all other agricultural holdings.

The advantage of using the indicator variable is that all estimates can be expressed as “national” level estimates. The indicator variable ${}_d\delta_i$ will automatically exclude those agricultural holdings that are not part of the estimation domain. It should be emphasized that the sampling weights were constructed based on the selection probability of the agricultural householders, which depends on the location (i.e. EA) of the householders whereas the estimation domain will be defined based on the attributes of the agricultural holding, e.g. location and type, etc.

4.3 Quality of the Survey Estimates

Because estimates are based on sample data, they will differ from figures that would have been obtained from complete enumeration of the population of agricultural

holdings using the same instrument. Results are subject to both non-sampling and sampling errors. Non-sampling errors include biases from inaccurate reporting, processing, and measurement, as well as errors from non-response and incomplete reporting. These types of errors cannot be measured readily. However, to the extent possible, each error can be minimized through the procedures used for data collection, editing, quality control, and non-response adjustment. The variances of the survey estimates are used to measure the sampling errors. The variance estimation methodology is discussed in this section.

4.3.1 Variance Estimation

The most commonly used methods for estimating variances of survey estimates from complex surveys, such as the UCA 2008/09, are the Taylor-series Linearization, Jackknife Replication, Balanced Repeated Replication (BRR), and Bootstrap methods (Wolter, 2007). We will use the JK2 Jackknife Replication method because of its simplicity. The JK2 Jackknife method is applicable when two primary sampling units (PSUs) are sampled from each stratum. Since we have selected EAs with PPS systematic sampling procedure by using a sort ordering that provides implicit stratification we can treat consecutive pairs of sampled EAs as 2 PSUs per stratum. In other words, by keeping the EAs in the order in which they appeared on the sample file, the first two EAs are paired to form a Jackknife (JK2) stratum; then the EAs 3 and 4 are paired; then the EAs 5 and 6 are paired, and so on. At the end of the process, $n/2$ Jackknife (JK2) strata would have been formed, each containing 2 EAs. Each pair is now treated as a stratum because of implicit stratification by sort ordering for the PPS systematic sampling.

Going over each JK2 stratum, one EA will be dropped at random, and the weights of the other EAs will be adjusted accordingly (i.e. doubling the base weight of the retained EA from the JK2 stratum, and repeating the post-stratification adjustment). The principle of the JK2 Jackknife method is to drop in turn one PSU (i.e. EA) at random, and re-compute the final weights to account for the loss of one EA, and produce an estimate of the characteristic of interest using this reduced sample. Thus, there are as many replicates as there are pairs of EAs in the full sample. The EAs will be paired within

Districts only, and if the number of EAs sampled from a District is an odd number then the last replicate in the District will be computed by dropping one complete EA at random, and then randomly dropping only half of the holdings from another EA which is also picked at random (i.e. dropping one and half EAs). The sampling variance is then estimated by computing the squared differences between each of the replicate estimates and the full sample estimate, and is given as:

$$v(\hat{\theta}) = \sum_{r=1}^R (\hat{\theta}_{(r)} - \hat{\theta})^2, \quad (4-4)$$

where

- θ is an arbitrary population parameter of interest;
- $\hat{\theta}$ is the estimate of θ based on the full sample;
- $\hat{\theta}_{(r)}$ is the estimate of θ based on the r^{th} JK2 replicate sample;
- R is the total number of JK2 replicates formed; and
- $v(\hat{\theta})$ is the estimated variance of $\hat{\theta}$.

It should be noted that the full sample is comprised of 3,612 EAs that would mean 1,806 replicate estimates and rather tedious computations. But there are collapsing techniques that can be used to reduce the number of replicates. For example, we can consider the first 4 EAs as one JK2 stratum, and collapse 1st EA with the 3rd EA to form a PSU and collapse 2nd EA with the 4th EA to form another PSU. Thus, the number of JK2 Jackknife replicates can be reduced from 1,806 to 903 by collapsing the EAs using the above scheme. Alternatively, for estimating variances of the estimated totals the replicates can be created for one district at a time to estimate the variances of the district level estimates of the totals. The variances of the national level estimates of the totals can then be obtained by aggregating over the districts.

The JK2 replication strategy described above is often adopted for its appealing simplicity. There are limitations, though, to what Jackknifing can accomplish. Jackknife is quite efficient at estimating variances for totals and functions of totals (e.g., ratios,

proportions). Jackknife is not as good when order statistics (e.g., percentiles) are of interest.

UBOS is currently using the software *STATA*³ for variance estimation. *STATA* provides the option of choosing among the Taylor-series Linearization, Bootstrap, Balance Repeated Replication (BRR), and Jackknife variance estimation methods. The stratified multi-stage sample designs are also supported by *STATA*. Moreover, the post-stratification feature is available in *STATA*.

Recommendation 10: We recommend that UBOS continue using the software *STATA* for variance estimation but may consider JK2 Jackknife method instead of Taylor-series Linearization method because of its simplicity. In either case, an interface would have to be built to create pseudo-strata and pseudo-PSUs as described above to reflect the efficiency gains from implicit stratification due to sort ordering by County and Sub-county for the PPS Systematic sampling. The gains due to post-stratification weighting are accounted for in the variance estimation.

4.3.2 Other Measures of Precision

In practice, the sampling variance is hardly ever reported. Instead, users find it more useful to rely on one of the derivatives of the sampling variance, such as the *standard error*, the *coefficient of variation*, the *margin of error*, or the *confidence interval*. These are all related expressions, and it is quite easy to go from one to the other using simple mathematical operations.

Standard Error

The standard error of an estimator is the square root of its sampling variance. This measure is easier to interpret since it provides an indication of sampling error using the same scale as the estimate whereas the variance is based on squared differences.

³ *STATA* is Data Analysis and Statistical Software developed by STATA CORP, Inc.

If $\hat{\theta}$ is the estimate of an arbitrary population parameter θ and $v(\hat{\theta})$ given in equation (4-4) is the corresponding estimate of its variance, then the standard of the estimate is defined as:

$$se(\hat{\theta}) = \sqrt{v(\hat{\theta})}. \quad (4-5)$$

Coefficient of Variation

It is more useful in many situations to assess the size of the standard error relative to the magnitude of the characteristic being measured. The ***coefficient of variation*** (cv) provides such a measure. It is the ***ratio of the standard error of the survey estimate to the value of the estimate itself expressed as percent***. It is very useful in comparing the precision of several different survey estimates, where their sizes or scale differ from one another. The coefficient of variation of $\hat{\theta}$ denoted by $cv(\hat{\theta})$ is defined as:

$$cv(\hat{\theta}) = 100 \times \frac{se(\hat{\theta})}{\hat{\theta}}. \quad (4-6)$$

Construction of Confidence Intervals

The 95 percent confidence interval is the interval such that there is a 95 percent probability (chance of 19 out of 20) of the unknown population parameter θ being within the interval. The 95 percent confidence interval is given by:

$$\hat{\theta} \pm 1.96 \times se(\hat{\theta}). \quad (4-7)$$

The lower limit of the interval is $\hat{\theta} - 1.96 \times se(\hat{\theta})$, and the upper limit of the interval is $\hat{\theta} + 1.96 \times se(\hat{\theta})$. The width $1.96 \times se(\hat{\theta})$ is known as half-width of the 95 percent confidence interval. The factor 1.96 is the z -value at $\alpha = 0.025$ for the standard normal

distribution. The factor 1.96 is often rounded to the approximate value 2.0. The smaller the half-width of the confidence interval, the more precise is the survey estimate.

Design Effects

Most surveys are based on complex designs involving stratification, and clustering due to multi-stage designs. Moreover, the weighting involves non-linear adjustments (e.g., non-response and post-stratification adjustments, etc.). It is crucial that these features of the complex survey design be accounted for in the variance estimation (Choudhry and Valliant, 2003). The ***design effect*** compares the variance of the estimate from the sample design that was actually implemented to the variance of the estimate that would have been obtained from an SRS design. ***Design Effect*** is another way to evaluate the efficiency of a sample design and the procedure used to develop the survey estimates. Design effect is defined as the ratio of the variance of an estimate for a complex sample design and the variance of the estimate under the simple random sample (SRS) design with the same sample size. Kish (1965) introduced the concept of design effect to deal with complex sample designs involving stratification and clustering. Stratification generally leads to a gain in efficiency over simple random sampling, but clustering leads to deterioration in the efficiency of the sample design due to positive intra-cluster correlation among units in the cluster (EA in this case). To determine the total effect of any complex design on the sampling variance in comparison to the alternative simple random sample design, the design effect (*deff*) is defined as:

$$Deff = \frac{\text{sampling variance of a complex sample design}}{\text{sampling variance of simple random sample design}}. \quad (4-8)$$

A design effect can be derived for any sampling design and estimator, provided we can compute a sampling variance. It is important to note that the design effect is associated with both the design and the estimator; therefore, for a given survey, the design effect can vary quite a lot from one variable to another.

Effective Sample Size

Another concept that is often used is *effective sample size* defined as the actual sample size that was selected for the complex design divided by the corresponding design effect. The effective sample size can be interpreted as the sample size that would be needed for an SRS design to obtain the same variance as that obtained with the complex design (i.e. the design that was actually implemented).

5. CONCLUSIONS AND RECOMMENDATIONS

We evaluated the sample design that UBOS was planning to implement for the UCA 2008/09 in September 2008. We have made a number of recommendations to improve the efficiency of the sample by taking into consideration the UBOS schedule and resource constraints. The following summarizes the main recommendations.

1. The information about *partnership agricultural holdings* needed for sample weighting must be collected. This can be collected on the Listing Module (UCA Form 1). Moreover, the questionnaire disposition codes (Result Codes) are also needed for sample weighting and must be collected as well.
2. QA checks must be implemented during all phases of the survey operation. In particular, the enumerator training must be very intensive and the supervisors must check their work thoroughly to improve quality. The Statistics Canada (2003) publication can be used as a guide to develop and implement QA procedures.
3. As part of the QA of the list frame, the respective supervisors in their districts must check a sample of agricultural businesses by field visit after the listings have been updated.
4. Power allocation with $\lambda = 0.4$ is recommended for allocating the area frame sample across districts. The only district that will have somewhat deficient sample is KALANGALA, the smallest size district and its sample will be increased to the same level as the sample for KAMULI, the 2nd smallest district.
5. We recommend that UBOS conduct the agricultural Core Module as a *piggy-back* with the upcoming Population and Housing Census in 2010. The Core Module should be conducted on a 100 percent basis.
6. UBOS should follow the WCA 2010 recommendations for the items to be included in the Core Module. The variable “Land Area of Holding” by Type of Use is an important auxiliary variable for designing future agriculture surveys and must included in the Core Module.

7. The area frame sample size should be changed to 3,612 EAs with sampling of 10 households per EA from the currently planned sample size of 3,200 EAs with sampling of 15 households per EA. Thus, the currently planned sample size of 48,000 households will be reduced to 36,120 households. The recommended sample of 36,120 households would result in an overall cost-variance efficiency of 111.4 percent relative to the currently planned sample because of reducing the clustering of sample from 15 households per EA to 10 households per EA.
8. If a “large” EA results in selection probability being greater than 1 then the EA should be treated as two or more pseudo EAs in a “conceptual split” such that each pseudo EA will have selection probability less than 1. On the other hand, the “very small” EAs (say, EAs with less than 20 agricultural households) should be collapsed with other EAs in the same neighborhood. It would be preferable to collapse the “very small” EA with a contiguous EA.
9. Because of the current budget situation we recommend that the Core Module be postponed until 2010 when it can be conducted on a 100 percent basis by piggy-backing onto the PHC 2010.
10. We recommend that UBOS continue using the software *STATA* for variance estimation but may consider JK2 Jackknife method instead of Taylor-series Linearization method because of its simplicity. In either case, an interface must be developed to create pseudo strata with two PSUs per stratum to reflect the stratification gains due to sort ordering the lists of EAs within the districts by County and Sub-county when selecting samples of EAs with the PPS systematic sampling procedure.

6. REFERENCES

Bankier, M.D. (1988), *Power Allocation: Determining Sample Sizes for Sub-national Areas*, *The American Statistician*, Vol. 42, No. 3, pp. 174-177.

Ching'anda, E.F. (2008), *Technical Report on a Consultancy on Methodological Studies for Agricultural Data (Area and Production)*, Prepared for DFID, February 2008

Choudhry, G.H., Lee, H., and Drew, J.D, (1985), *Cost-Variance Optimization for the Canadian Labor Force Survey*, *Survey Methodology*, 11, pp. 33-50.

Choudhry, G.H. and Valliant, R. (2003), *WesVar: Software for Complex Survey Data Analysis*, *Proceedings of the Statistics Canada Symposium on Analysis of Data from Complex Surveys*, Ottawa, Ontario, Canada.

Food and Agriculture Organization of the United Nations (2005), *World Programme for the Census of Agriculture 2010: A System of Integrated Agricultural Censuses and Surveys Volume 1*, *FAO Statistical Development Series 11*.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons

Kish, L. (1965), *Survey Sampling*, John Wiley & Sons

Kish, L. (1988), *Multi-purpose Sample Designs*, *Survey Methodology Journal*, Vol. 14, pp. 19-32.

Statistics Canada (2003), *Statistic Canada Quality Guidelines Fourth Edition*, *Statistics Canada Catalogue No. 12-539-XIE*, October 2003.

Uganda Bureau of Statistics (2004a), *Report of the Pilot Census of Agriculture (PCA) 2003*, UBOS Report, February 2004.

Uganda Bureau of Statistics (2004b), *Report on the Agricultural Module, Piggy-Baked onto the Population and Housing Census (PHC) 2002*, UBOS Report, September 2004.

Uganda Bureau of Statistics (2005), *2002 Uganda Population and Housing Census*, UBOS Main Report, March 2005.

Uganda Bureau of Statistics (2006), *2002 Uganda Population and Housing Census, Analytical Report (Abridged Version)*, October 2006.

Uganda Bureau of Statistics (2007a), *Uganda National Household Survey 2005/2006, Report on the Agricultural Module*, UBOS, April 2007.

Uganda Bureau of Statistics (2007b), *Report on the Uganda Business Register 2006/2007*, UBOS, June 2007.

Uganda Bureau of Statistics (2007c), *Uganda Census of Agriculture 2008/2009*, Project Document, UBOS, September 2007.

Uganda Bureau of Statistics (2008b), *Questionnaires and Manuals being developed for the Uganda Census of Agriculture 2008/2009*

United Nations (2003), *Indicators for Monitoring the Millennium Development Goals – Definitions, Rationale, Concepts and Sources*, New York

Westat (2002), *WesVar Version 4*, Rockville, Maryland, USA

Wolter, K. M. (2007), *Introduction to Variance Estimation, 2nd Edition*, Springer-Verlag: New York.

ANNEXES

ANNEX 1

Terms of Reference for the General Data Dissemination System (GDDS), Phase 2

Statistics Projects for Anglophone Africa

Provision of technical assistance as the expert for Agriculture Statistics for Uganda, Uganda Bureau of Statistics (UBOS)

Background

With financial support from the Department for International Development (DFID) of the United Kingdom, the World Bank is implementing a project to assist 21 Anglophone Africa countries to participate in the General Data Dissemination System (GDDS). Participating countries are being assisted to participate in the GDDS through two separate, but linked projects both financed by DFID. The IMF is providing project management and technical support in the area of economic and financial statistics. The World Bank is providing technical support in the area of socio-demographic statistics. Both projects run concurrently until March 2009.

Technical Assistance

Technical assistance is being provided through the World Bank to help countries implement plans for improvement in population, health, agriculture, labor market, justice and security, management of statistical systems, GIS and small area statistics. The GDDS framework developed by the IMF provides the framework for the detailed elaboration of long-term statistical development strategies. Participating countries have already expressed their requests for technical assistance and both the IMF and the World Bank have developed their assistance strategies. **Uganda** was one of the countries which asked for technical assistance in the field of Agriculture Statistics.

Uganda attended the GDDS 2 Module launch workshop on Agriculture statistics in **Maputo in March 2007**, where they drew up their Country Work Plan regarding the deliverance of three Technical Assistance Missions covering the three priorities that the country had identified. These priorities are part of the Work Plan Structure Document.

The purpose of the work plan structure document is to act as a living document for the duration of the technical assistance and to serve as an information base from which the terms of reference (TOR) for every mission can be drawn up. To this end, this TOR for the first mission to Uganda has been drawn up from the work plan.

UBOS's general objective regarding the topic is to have GDDS assistance to design a sample for the census of agriculture. Other missions will be carried out by other experts at other times.

Purpose of the Assignment

The purpose of the assignment is to complete the first technical assistance mission aimed at helping UBOS establish an appropriate sampling design for their Agricultural Census. The major activities for this assignment for this mission will include:

- Formulation of an appropriate sampling design for the Agricultural Census,
- Drafting a mission report for the country file covering the topics discussed.

The consultant will be asked to provide assistance to UBOS using the following outline as a guide:

Activities

In providing assistance to UBOS in designing an appropriate sampling design for the Agricultural Census, the consultant will take the following factors into account:

- The objectives and the major uses intended for the outputs from the Census of Agriculture.
- The characteristics of the Census of Population which will serve as the frame for drawing the sample for the Census of Agriculture.
- The availability of supplementary lists of agricultural operations to be used to supplement the census of population frame, e.g. large flower greenhouse operation south of Kampala. If a list of larger operations does not exist, suggestions should be made to assist UBOS in establishing such a tool.
- The budgets available and the priorities for the Census of Agriculture.

- Sample design, size/structure/stratification taking into account impact on field operations and budgets. Sampling procedures should be discussed with UBOS officials.
- Assistance to UBOS staff in selecting and finalizing a sample of agriculture units should be given with the aim of instructing them on sampling procedures.
- Measurement issues, especially for crop land. While a complete review of the proposed agriculture questionnaire may be beyond the scope of this mission, suggestions with regard to feasibility of questions should be provided. A strategy for testing questions should be discussed.
- Weighting and estimation procedures.

Skill requirements

This mission will require an expert who can read and write English fluently and is knowledgeable in the design and implementation of sample household/agriculture surveys in an African context.

Duration

The total consultant time for this mission is 13 working days with 10 days mission time and 3 days preparation and reporting time.

Timing

The field work for this mission is to be carried out during the time period from March 31 to April 11, 2008. The final report is to be completed by April 30, 2008.

Duration of the consultancy: 13 working days

Starting date: March 31, 2008

Location: Kampala, Uganda

ANNEX 2

List of Officials Met and Meeting Notes

Meeting with Mr. SETH N. MAYINZA, Director, Division of Agriculture Statistics, UBOS on March 31, 2008

This was a briefing meeting with Mr. MAYINZA to discuss the program for the consultancy mission. Mr. MAYINZA also provided the background documents for the UCA 2008/09 project.

Meeting with the UCA 2008/09 Project Team on April 1, 2008

Present

Mr. SETH N. MAYINZA, Director, Division of Agriculture Statistics, UBOS

Mr. MUWDNGE JAMES, Senior Statistician, Division of Population and Social Statistics, UBOS

Mr. PATRICK OKELLO, Senior Statistician, Division of Agriculture Statistics, UBOS

Mr. MENYHA EMMANUEL, Senior Statistician, Division of Agriculture Statistics, UBOS

Mr. NSIKO ISRAEL, Statistician, Division of Agriculture Statistics, UBOS

Discussed the following issues related to the proposed sample design:

- 1) Construction of a list frame for large units,
- 2) Sample allocation across districts – compromise allocation,
- 3) Average number of households to be sampled per EA,
- 4) Sample size for the Core Module.

The need for training the UBOS technical staff in sample design and weighting was also discussed, and it was decided to conduct three training sessions to provide training in survey sampling.

Meeting with Mr. BYLON TWESIGYE, Field Operations Officer, Division of Socio-Economic Surveys, UBOS and Mr. SSENONO VINCENT, Senior Statistician, Division of Population Social Statistics, UBOS

April 4, 2008

The number of districts in the country was increased from 56 to 80 in 2005 by splitting 13 old districts into 2 each, 4 old districts into districts into 3 each, and one old district into 4 new districts. The 2008/2009 UCA will be designed to produce estimates for the new district boundaries. The data collected through the Agriculture Module in the 2002 PHC is for the 56 districts, and Agriculture Division has not yet converted the data files from the old districts to the new districts. The purpose of the meeting with **Mr. BYLON and Mr. VINCENT** was to get the counts of Agricultural Households for the new districts by matching the file from the Agriculture Division with the file that provides the composition of the new districts in terms of counties and sub-counties. The composition of the new districts is given in terms of county and sub-county names only, and there was no Standard Geographic Codes (SGC) on the files, and these had to be matched by name. The problem with matching the counties and sub-counties by name was there were a number of non-matches because of slight differences in the names. The non-matched counties and sub-counties had to be resolved manually. **Mr. VINCENT** did the painstaking job of checking all the non-matched cases, and manually coded them to the numeric codes. He was able to match 100 percent of counties and sub-counties, and he created a file of the new districts with the count of agricultural households.

Meeting with Mr. KYEWALYANGA SIMON, Statistician, Division of Population and Social Statistics, UBOS

April 7, 2008

Mr. Simon is responsible for computing variances of the survey estimates produced from UBOS household and agricultural surveys. The purpose of the meeting was to find out about the software that UBOS is using, and the methodology employed for variance estimation. UBOS is using the software *STATA* for variance estimation employing Taylor-series Linearization method. This is quite satisfactory as the software *STATA* takes into account the complex survey design when estimating variances of the estimates.

Meeting with Mr. J.B. MAGEZI-APUULI, Principal Statistician, Directorate of Agriculture Statistics, UBOS on April 9, 2008

The purpose of the meeting was to discuss the budget implication of recommended sample size of 3,612 EAs with 10 households per EA instead of the current plan of selecting 3,200 EAs with 15 households per EA. Mr. MAGEZI-APUULI had already estimated the field cost of the current plan of sampling 3,200 EAs with 15 households per EA, which was USH 16.639B. Mr. MAGEZI-APUULI estimated the field cost for the recommended alternative to be USH 15.753B, a saving of 5.3 percent.

Meeting with the 2008/2009 UCA Project Team on April 9, 2008

Present

Mr. MUWDNGE JAMES, Senior Statistician, DPSS, UBOS

Mr. PATRICK OKELLO, Senior Statistician, Division of Agriculture Statistics, UBOS

Mr. MENYHA EMMANUEL, Senior Statistician, Division of Agriculture Statistics, UBOS

The meeting was called to discuss the sample allocation across districts using Power Allocation. Implementation of the PPS systematic sampling algorithm to select the sample of EAS was also discussed.

Meeting with Professor E.S.K. MUWANGA-ZAKE, Institute of Statistics and Applied Economics, Makerere University, Kampala on April 9, 2008

Professor MUWANGA-ZAKE is a local consultant and a former UBOS employee, and will be working on the UCA 2008/09 project under a contract with UBOS. Professor MUWANGA-ZAKE was interested in discussing my recommendations in particular, how lower variance was achieved with 25 percent smaller sample.

Meeting with Mr. PATRICK OKELLO, Senior Statistician, Division of Agriculture Statistics, UBOS and Mr. MENYHA EMMANUEL, Senior Statistician, Division of Agriculture Statistics, UBOS on April 10, 2008

The purpose of the meeting was to assist UBOS statisticians to implement the SAS code to select the sample of EAs with PPS systematic sampling procedure. The SAS code was implemented and tested by selecting sample of EAs from two districts.

Briefing Meeting with the UBOS Senior Management on April 11, 2008

The purpose of the meeting was to brief the UBOS senior management about the changes recommended for the UCA 2008/09. The meeting started with a power point presentation highlighting the impact on the variance and the cost of the recommended alternative. Both the variance and cost are lower for the recommended alternative, and the overall cost-variance efficiency of the alternative is 111.4 percent as compared to the sample that had been planned. The project budget and schedule were also discussed. The UBOS project budget was actually in deficit and the recommended alternative turned the deficit budget into a balanced budget. Therefore, decision was taken to adopt the recommended alternative. The implications of the tight schedule for the development and implementation of QA procedures were also discussed. Finally, the implications of not following the WCA 2010 recommendations for the Core Module were discussed as well.

ANNEX 3

Sample Allocation across District under Power Allocation ($\lambda = 0.4$)

District Name	Agriculture Households	Sample Size (EAs)
KALANGALA	3,508	23
KAMULI	7,827	23
BUKWO	8,371	23
ABIM	8,772	24
BULLISA	9,453	25
LYANTONDE	10,757	26
KOBOKO	14,732	29
AMOLATAR	16,300	30
KAMPALA	17,560	31
KOTIDO	19,526	33
NAKASONGOLA	22,000	34
BUDUDA	22,686	35
NAKAPIRIPIRIT	23,339	35
BUKEDEA	23,422	35
KABERAMAIDO	23,846	35
BUDAKA	23,888	36
DOKOLO	24,314	36
KAPCHORWA	24,401	36
KALIRO	26,580	37
KATAKWI	28,393	38
BUTALEJA	28,805	38
NAMUTUMBA	29,251	39
MOROTO	30,363	39
AMURIA	30,545	39
AMURU	32,874	40
KIRUHURA	33,184	41
SSEMBABULE	33,339	41
ADJUMANI	33,524	41
MOYO	34,016	41
BUSIA	35,229	41
IBANDA	36,390	42
BUNDIBUGYO	37,144	42
KANUNGU	38,911	43
GULU	39,140	43
YUMBE	39,144	43
KITGUM	40,530	44
KISORO	42,090	45
KIBOGA	42,265	45
MBALE	42,486	45
KAABONG	43,165	45

MITYANA	45,060	46
MAYUGE	47,561	47
JINJA	48,443	47
KAYUNGA	48,652	47
KUMI	49,373	47
MANAFWA	49,409	47
RUKUNGIRI	49,484	48
OYAM	49,644	48
KAMWENGE	52,108	49
MBARARA	53,526	49
NYADRI	54,615	49
HOIMA	55,905	50
ISINGIRO	57,290	50
SOROTI	57,481	50
PADER	57,865	51
KABAROLE	59,524	51
SIRONKO	59,538	51
MASINDI	61,265	52
RAKAI	61,727	52
ARUA	62,047	52
PALLISA	66,203	53
NTUNGAMO	67,442	54
BUGIRI	68,032	54
TORORO	68,502	54
MPIGI	69,893	55
KYENJOJO	70,371	55
WAKISO	75,146	56
KIBAALE	75,737	56
MUBENDE	78,205	57
KASESE	78,285	57
APAC	78,540	57
NEBBI	80,082	58
LUWEERO	82,491	58
IGANGA	83,506	59
KABALE	83,541	59
LIRA	88,617	60
KAMULI	95,132	62
MUKONO	113,041	66
MASAKA	123,738	69
BUSHENYI	124,394	69
National Total	3,833,485	3,612

ANNEX 4

Probability Proportional to Size (PPS) Systematic Sampling Procedure

In this Annex we describe the probability proportional to size (*PPS*) systematic sampling procedure for selecting EAs from within districts. For the sake of simplicity we will not use any subscript to denote the district. However, it is understood that the procedure will be applied independently within each district.

Let N be the total number of EAs in the district, and the number of EAs to be selected from the district is denoted by n . Also, let z_i be the size measure of the EA labeled i within the district, where $i = 1, 2, \dots, N$.

Define the total of the measure of size (MOS) for the district as:

$$Z = \sum_{i=1}^N z_i.$$

A procedure for selecting a sample of n EAs out of the N EAs in the district with *PPS* systematic sampling procedure can be implemented as follows.

Step 1: Sort the list of EAs in the district by counties and sub-counties. Thus, the sub-counties will be the implicit strata. Next, sort the block of EAs within the 1st implicit stratum by MOS in ascending order followed by the block of EAs within the 2nd implicit stratum by MOS in descending order and so on, **by alternating between ascending and descending sort** from one implicit stratum to the next.

Step 2: Check that z_i is less than Z/n for all i in the district.

Step 3: Define the relative size measures p_i as:

$$p_i = \frac{z_i}{Z}; i = 1, 2, 3, \dots, N.$$

Also, define the probabilities of selection π_i as:

$$\pi_i = n \times p_i; i = 1, 2, 3, \dots, N.$$

The π_i values will be called “**Normalized Size Measures**”.

Step 4: Compute cumulative totals $C_1, C_2, C_3, \dots, C_N$ as:

$$C_1 = \pi_1$$

$$C_2 = C_1 + \pi_2$$

$$C_3 = C_2 + \pi_3$$

$$C_4 = C_3 + \pi_4$$

.

.

$$C_{N-1} = C_{N-2} + \pi_{N-1}$$

$$C_N = C_{N-1} + \pi_N.$$

Note that the cumulative total C_N must be equal to n , the number of EAs to be selected from the district.

Step 5: Generate a random number " r " from the uniform distribution between 0 and 1, and compute n numbers $r_1, r_2, r_3, \dots, r_n$ as follows.

$$r_1 = r$$

$$r_2 = r + 1$$

$$r_3 = r + 2$$

.

.

$$r_i = r + i - 1$$

.

.

$$r_n = r + n - 1$$

Step 6: Select the n EAs out of the N EAs in the district with the labels $i_1, i_2, i_3, \dots, i_n$ such that

$$C_{i_1-1} < r_1 \leq C_{i_1}$$

$$C_{i_2-1} < r_2 \leq C_{i_2}$$

$$C_{i_3-1} < r_3 \leq C_{i_3}$$

.

.

$$C_{i_n-1} < r_n \leq C_{i_n}$$

The above n EAs would then get selected with probabilities proportional to size, and the selection probability of the EA labelled i within the district will be given by $\pi_i = np_i$, where n is the number of EAs to be selected from the district.