Technical Note No. 1

Sample and Survey Methodology

SAMPLE AND SURVEY METHODOLOGY

Background:

This section describes the statistical procedures adopted for the selection of the Yemen National Social Protection and Monitoring Survey (NSPMS) household sample. Survey samples for impact evaluation and monitoring of welfare programmes need to be thought in the context of providing comparative assessment between population subgroups of programme beneficiaries and non-beneficiaries along several points in time. The target population of the NSPMS is the Yemeni resident population (excluding non-household communities such as refugees, nomads and internally displaced persons, hotels, dormitories, prisons and hospitals). NSPMS is a longitudinal household survey that will last for 12 months and aims to provide parameter estimates quarterly as well as to accommodate the Social Welfare Fund (SWF) programme impact assessment at the round 4.

Yemen is geographically organised into 21 governorates, divided into 333 districts. Districts are subdivided into sub-districts, and each sub-district further subdivided into villages. The Yemeni census data collection system defines enumeration areas (EAs) within districts. Initial reports indicated that each EA should cover approximately 130 households in rural, and between 150 and 180 households in urban areas. Nevertheless, it has been noticed that there are currently cases of EAs that have either fewer than 130 households in rural areas or more than 180 households in urban areas.

Sample design:

The NSPMS sample is selected following a stratified two-phase sampling design. In Phase 1, EAs are considered as primary sampling units, and a stratified cluster sampling procedure is implemented with unequal selection probabilities, considering governorates as strata. In Phase 2, households are selected from each selected EA by a stratified simple random sampling procedure. Detailed descriptions of each sampling phase are provided in subsections a) and b) below.

The total survey sample size was previously set to approximately 7560 households. In each governorate, 30 EAs are selected, and in each sampled EA, 12 households are selected. Sample size allocation to Phase 2 strata took into account post-survey analysis needs, focusing on providing higher probabilities of finding counterfactual matches for the treatment cases in the policy impact econometric analysis to be performed.

a) *The Phase 1 sample*

In the first phase of the NSPMS sample design, EAs are geographically stratified by governorate and selected using a probability proportional to size (PPS) sampling scheme, following a sequential Poisson sampling procedure (Ohlsson, 1998).

Let μ_{hk} be the average number of poor people per EA in district *k*, and governorate *h*, with

$$\mu_{hk} = \frac{\tau_{hk}}{M_{hk}},$$

where M_{hk} and τ_{hk}denote the number of EAs and the total number of poor people in district *k* , of governorate *h*, respectively, obtained from the Household Budget Survey 2005/6 district-level data. In the sampling algorithm described below, for every EA in district k (and governorate h), μ_{hk} is taken as the size variable in the application of the sequential Poisson sampling procedure. Given a governorate *h*, the following algorithm is applied to select the EA sample:

Step I. For each EA *i* in the given governorate *h*, generate u_i from a uniform(0,1) distribution;

Step II. Calculate $p_{hi} = \mu_{hk} \Big/ \sum_{k=1}^{M_h} \mu_{hk}$ for each EA *i*;

Step III. Calculate $Z_{hi} = u_i / p_{hi}$ for each EA *i*;

Step IV. After ordering the listed EAs according to increasing values of Z_{hi}, select the first 30 to the sample.

It should be noted that EAs within the same district have the same size variable value and, hence, the same selection probability. Analysis of the resulting size measures, before implementing the sampling algorithm, revealed no cases with extreme values. In particular there were no EA with zero, or close to zero average number of poor people assigned.

Let π_{hi} and γ_{hi} be the first-phase selection probability and the size variable for the EA *i* within governorate *h*, respectively. Hence, $\gamma_{hi} = \mu_{hk}$ for every EA *i* within district *k*. Also let m_h be the number of EAs selected in governorate *h*. Therefore,

$$\pi_{hi} = m_h \frac{\gamma_{hi}}{\sum_{i \in U_{lh}} \gamma_{hi}}$$

where U_{lh} denotes the frame set of EAs listed in governorate *h*. Sampling with probability proportional to a poverty measure is expected to increase statistical inference efficiency and also the efficiency in targeting the population more likely to participate in the SWF programme.

Designing samples to assess welfare programmes requires survey statisticians to focus on the estimation of parameters related to comparative studies. Phase 1 strata sample sizes were allocated

based only on the post-survey analysis criterion of producing contrast estimates for each stratum, with similar levels of precision. By the time the first-phase sample was designed, estimation of a contrast between sub-populations of beneficiary (*b*) and non-beneficiary (*nb*) households was the main goal. Following Cochran (1977: section 5A.13), ignoring finite population correction factors, optimum allocation rules for simple random samples in each stratum can be derived by minimising

$$Var(\bar{y}_b - \bar{y}_{nb}) = \frac{\sigma_b^2}{n_b} + \frac{\sigma_{nb}^2}{n_{nb}}$$

subject to cost constraints such as

$$C = c_0 + c_b n_b + c_{nb} n_{nb}.$$

The results indicate that minimisation of the variance is achieved when

$$n_b = n \frac{\sigma_b / c_b}{\sqrt{\sigma_b^2 / c_b + \sigma_{nb}^2 / c_{nb}}}$$

and

$$n_{nb} = n \frac{\sigma_{nb} / c_{nb}}{\sqrt{\sigma_b^2 / c_b + \sigma_{nb}^2 / c_{nb}}}.$$

With no further information on variances, and assuming approximately equal costs of sampling, allocation assuming equal variances ($\sigma_b^2 = \sigma_{nb}^2$) and costs ($c_b = c_{nb}$) can be done uniformly through the sub-populations of interest. It should be noted that, differently from the usual sampling scenario, when dealing with contrast (counterfactual) estimation for impact evaluation studies, population sizes have no effect on the allocation decision. These known theoretical results support the decision to adopt uniform sample allocation of 30 EAs per stratum for the NSPMS and explain why population size was not considered a major criterion for sample allocation, as the precision of estimators is mainly affected by the variability of the target variables.

b) The Phase 2 sample

The second phase of the NSPMS sample is designed based on a screening (listing) operation conducted at each selected EA at the first phase of the sampling procedure. A screening process was carried out between 1st and 22nd September 2012 with the aim of identifying and classifying every household located in each of the selected EAs by collecting the information requested on the listing questionnaire (see Annex).

Within each EA selected in Phase 1, the Phase 2 sample design is stratified by the household classification obtained from questions (q7), (q8) and (q9) of the listing questionnaire. The following

three strata were formed: (*i*) treatment stratum: households with at least one beneficiary of the SWF programme (with at least one payment already received); (*ii*) control 1 stratum: households with at least one resident either selected or registered for the SWF programme but without any beneficiaries; and, (*iii*) control 2 stratum: households with all their residents not registered for the SWF programme.

Moreover, a simple random sample of households is selected in each stratum, which characterises a stratified simple random sampling design in the second phase. It should be noted that such a design, implemented within each EA selected in the first phase, preserves both invariance and independence properties (Särndal, Swensson and Wretmann, 1992: 134), which is desirable, as it facilitates field work.

In the second-phase sampling, the 12 households selected from each (first-phase) sampled (and screened) EA are allocated into the classification strata as indicated in Box M. 1.

Box M. 1: **Second-phase sample allocation to classification strata**

Stratum	Description	Sample allocation
i)	Treatment	5
ii)	Control 1	5
iii)	Control 2	2

As can be seen in Box M.1, five households are allocated to the treatment stratum, five to the control 1 stratum, and two to the control 2 stratum. Such allocation aims to increase the probability of finding counter-factual matches for the treatment cases in the impact analysis.

In cases where the population size was smaller than the sample size allocated to any of the classification strata groups, the following protocol was adopted: (*i*) if the population size was smaller than five in the treatment group, the missing sample size was allocated to the control 1 group or, if that was not possible, to the control 2 group; (*ii*) if the population size was smaller than five in the control 1 group, the missing sample size was allocated to the control 2 group or, if that was not possible, to the treatment group; and (*iii*) if the population size was smaller than two in the control 2 group, the missing sample size was allocated to the control 1 group or, if that was not possible, to the treatment group.

Let $\pi_{j|hi}$ be the selection probability of a household *j* within the first-phase selected EA *i* in governorate *h*. Let $n_{g|hi}$ and $N_{g|hi}$ be the sample size and the total number of households, respectively, in the classification stratum *g* within the first-phase selected EA *i* of governorate *h.* Therefore,

$$\pi_{j|hi} = \frac{n_{g|hi}}{N_{g|hi}} \forall j \in U_{g|hi},$$

Where$\mathrm{U_{g|hi}}$ denotes the listing set of households from the first-phase selected EA *i* of governorate *h*, classified into stratum *g*.

According to the NSPMS sample design, the first-order inclusion probability of household *j* in EA *i* within governorate *h* is given by

$$\pi_{hij} = \pi_{hi}\pi_{j|hi}.$$

Longitudinal design

The NSPMS longitudinal data collection process follows a multiple panel rotation scheme, as illustrated in Box M.2. Two competing strategies for composing the panels are introduced in subsections a) and b).

a) *Panel composition strategy 1*

In Strategy 1 each panel is composed of a group of seven governorates, which could be chosen by the survey team to facilitate mobility of the data collection team.

Box M. 2: Multiple-panel rotation scheme

Panels	Months											
	1	2	3	4	5	6	7	8	9	10	11	12
Panel 1	1			2			3			4		
Panel 2		1			2			3			4	
Panel 3			1			2			3			4

As shown by Box M.2, the selected EAs located at governorates allocated to Panel 1 should be visited for the first time in the first month of the survey, for the second time in the fourth month of the survey and so on, totalling four visits. Panel 1's last visit is in the 10th month. It should be noted that, based on this strategy, the country data collection process is spread over a three-month period. This allows national estimates to be released quarterly, as expected, and policy impact to be monitored while the survey is being conducted. Strategy 1 also allows the production of monthly estimates at the governorate level, for those governorates included in each of the monthly panels.

Box M.3 shows how the sample sizes (number of EAs in the sample) are allocated in a given panel, for the first phase of the sampling process, according to Strategy 1. In a given month, 210 EAs are covered across the seven governorates within a panel. At the end of each three-month period, a total of 630

EAs are visited throughout the country.

Box M. 3: First-phase sample size (enumeration areas) longitudinal allocation using the adopted multiple-panel rotation scheme (Strategy 1)

Governorate	Month 1	Month 2	Month 3	Total
1	30			30
2	30			30
3	30			30
4	30			30
5	30			30
6	30			30
7	30			30
Sub-total	210			
8		30		30
9		30		30
10		30		30
11		30		30
12		30		30
13		30		30
14		30		30
Sub-total		210		
15			30	30
16			30	30
17			30	30
18			30	30
19			30	30
20			30	30
21			30	30
Sub-total			210	630

b) *Panel composition strategy 2*

In Strategy 2 the multiple-panel rotation scheme illustrated in Box M.3 remains valid. Each panel, however, is now considered to be composed of the 21 Yemeni governorates. In each month, 10 randomly allocated EAs, from the 30 sampled EAs at each governorate, are visited.

As shown in Box M.3, the 210 EAs selected and randomly allocated to Panel 1 should be visited for the first time in the first month of the survey, for the second time in the fourth month of the survey and so on, also totalling four visits. Box M.4 shows how the sample sizes are allocated for each panel for the first phase of the sampling process according to Strategy 2. In a given month, one third of the national sample (210 EAs) is covered across all the 21 governorates, with a total of 630 EAs visited in

the whole country every three months[1].

Box M. 4:**First-phase sample size (enumeration areas) longitudinal allocation using the adopted multiple-panel rotation scheme (Strategy 2)**

Governorate	Month 1	Month 2	Month 3	Total
1	10	10	10	**30**
2	10	10	10	**30**
3	10	10	10	**30**
4	10	10	10	**30**
5	10	10	10	**30**
6	10	10	10	**30**
7	10	10	10	**30**
8	10	10	10	**30**
9	10	10	10	**30**
10	10	10	10	**30**
11	10	10	10	**30**
12	10	10	10	**30**
13	10	10	10	**30**
14	10	10	10	**30**
15	10	10	10	**30**
16	10	10	10	**30**
17	10	10	10	**30**
18	10	10	10	**30**
19	10	10	10	**30**
20	10	10	10	**30**
21	10	10	10	**30**
Total	**210**	**210**	**210**	**630**

It should be noted that Strategy 2 allows national estimates to be released monthly at reduced precision levels, and quarterly with higher precision. Moreover, it only allows the production of estimates at the governorate level quarterly.

The NSPMS longitudinal strategy

To accommodate field work mobility restrictions and ensure better data collection quality, Strategy 2 was adopted as the NSPMS longitudinal scheme, as recommended by the local technical committee. Such a design can accommodate the monitoring of the necessary indicators at four points in time, generating estimates for the level of and change in the desired parameters. It will also allow impact estimation when comparing between the last quarter (round 4) and the first quarter baseline (round 1) parameter estimates.

[1] One can see in Map 1 in the annex the sub-districts for which there is at least one Enumeration Area (EA) in the sample. It is clear from the map that the sample is distributed all over the country.

Non-response handling protocol

Although several efforts are made to ensure completeness of responses, every survey is susceptible to non-response units and items. For the NSPMS survey, it is possible to identify at least three occasions when non-response has to be handled: in the first and the second phases of the sampling procedure and also during the follow-up waves of the longitudinal study.

Handling non-response in the first phase of sampling:

The first phase of sampling selects the EAs to be screened in each governorate. Such EAs are clusters of households to be further subject to a second sampling phase. Whenever the fieldwork team could not access a selected EA for security reasons, these EAs were replaced by new ones. Moreover, in some cases the selected EAs were later identified as not being part of the target population – such as a nomadic community or desert region. In these cases, the EAs were also replaced. In any case, the replaced EA is considered out of the target population. The impact of these replacements is expected to be negligible, as indeed some of the EAs are not included in the target population, by definition, and the remaining cases are expected to be very limited in number.

Handling non-response in the second phase of sampling:

The second phase of sampling deals with the selection of households where the measurement instrument (questionnaire) is applied. On the one hand, the occurrence of units of non-response is mainly handled by re-weighting methods of estimation, according to an analysis of the likelihood of the non-response generation mechanism. On the other hand, item non-response cases are mainly handled by imputation techniques suitable for each specific case.

Handling non-response during the longitudinal study:

Attrition and difficulties in following previously selected household families moving across regions are expected to be the main sources of unit and item non-responses during the longitudinal study. Re-weighting, regression and imputation methods are expected to be used as tools for handling these cases.

Sampling weights:

Cross-sectional sampling weights described in this section reflect not only the NSPMS sampling design but also the application of adjustment terms for dealing with unit non-response cases found in the first wave (i.e. first quarter) of the data collection process. Retaining notation used by Vieira and Ferraz (2013), let the following quantities be defined:

- π_{hi} is the first-phase inclusion probability of EA *i* within governorate *h* ;
- $\pi_{gj|hi}$ is the second-phase conditional inclusion probability of household j within group *g* given the selected EA *i* within governorate *h*;
- $\pi_{higj} = \pi_{hi}\pi_{(gj|hi)}$ is the inclusion probability of household j within group *g* , at EA *i* of governorate *h* .

The starting point for building the sampling weights is the Horvitz-Thompson (1952) estimator for a population total, given by

$$\hat{Y} = \sum_h \sum_i \sum_g \sum_j d_{higj} y_{higj}$$

,

where

- $d_{higj} = 1/\pi_{higj}$ is the basic sampling design weight for household j within group *g* , at EA i of governorate *h*; and
- y_{higj} is the value of the interest variable y for household j within group *g*, at EA *i* of governorate *h*.

In this simpler form, the basic sampling weights $\mathrm{d_{higj}}$ reflect only the main aspects of sampling design – i.e. the inverse of selection probabilities. For instance, since the first-phase sampling considers selection with probability proportional to a poverty measure, the estimation process assigns higher sampling weights to those EAs with lower poverty levels, and lower sampling weights to those with higher poverty levels.

The first-wave data revealed that, as expected in any survey, unit non-response cases were observed. Such cases were assumed to be generated by a missing at random (MAR) mechanism (Rubin, 1976), where the missing pattern depends only on the stratification sampling design variables (i.e. governorate and group). To cope with this problem, a weighting adjustment procedure was adopted to correct the basic sampling design weight.

Let

$$\hat{\phi}_{hig} = \frac{m_{h(r)}}{m_h} \frac{{}^{ng}/_{hi(r)}}{{}^{ng}/_{hi}}$$

be the estimated propensity score for responses within group *g* at EA i within governorate h. In this expression, m_h is defined as the number of EAs selected in governorate h, $m_{h(r)}$ is the number of EAs

that were actually surveyed, $^{ng}/_{hi}$ is the sample size in the classification stratum *g* within EA i of governorate h, and $^{ng}/_{hi(r)}$ is the number of households that actually responded to the questionnaire in the classification stratum *g* within EA i of governorate h.

Then:

- $\widetilde{w}_{higj} = \frac{d_{higj}}{\hat{\phi}_{hig}}$ is the sample weight adjusted for unit non-response at governorate h, within EA *i* and group g.

The household adjusted sample weight $\widetilde{w}$ is still subject to the effect of some households with extreme weight values that could cause variance inflation. To investigate this undesired effect in the households' non-response adjusted sampling weights, the adoption of a weight-trimming procedure was evaluated following methodology discussed by Potter (1988: Section 2). The trimming method is based on evaluations of the effect of trimming on the estimates of variables of interest means and sampling variances. Two variables of interest were investigated: total income and total of rural population. After the application of this methodology, the indicated weight-trimming value for controlling total income was the 94th percentile of the original non-response adjusted sampling weights. Let $q_{0.94}$ represent the 94th percentile of the original non-response adjusted sampling weights, r_{higj} be the adjustment term for non-extreme original weights, and r'_{higj} be the adjustment term for the extreme original weights. Therefore, the suggested final sampling weight w_{higj} would be expressed as:

$$w_{higj} = \begin{cases} w_{higj} r_{higj}, & if \quad \tilde{w}_{higj} < q_{0.94} \\ q_{0.94} r'_{higj} & if \quad \tilde{w}_{higj} \geq q_{0.94} \end{cases}$$

However, when analysing results for controlling the total of rural population, no trimming adjustment was recommended. Since both variables are of interest, the final sampling weights do not receive any further adjusting procedure, keeping its expression as introduced before. Hence,

$$w_{higj} = \widetilde{w}_{higj}$$

Item non-response

Although the cross-sectional sampling weight introduced in Section 3 takes into account unit non-responses for the first-wave data, item non-responses eventually found in the survey data set should be handled with additional care.

Imputation of missing item values has the advantage of keeping the information collected for those that responded. Several imputation techniques are available in the literature. Hot deck techniques consist of taking one of the observed values of the sample as the imputed one. In its simpler form, a

partition of the survey data set is determined by cross-classifying observations into two or more variables that are available and define the characteristics of the missing data. Within each cell of the partition, a donor is randomly taken to furnish the response value as the imputed information, when necessary.

When using imputation to cope with item non-response, it is advisable to keep an indicator variable for the imputed values, as they are artificially generated. This information can be useful for estimating appropriate variances and keeping track of the original data set.

Sample Coverage:

Table M.1 presents the number and percentage of interviewed households and individuals in the baseline survey (fieldwork conducted in October, November and December 2012). Due to security reasons the Saa'da sample was not interviewed. Of the 7,152 households selected for the sample, 6,969 were successfully interviewed and kept for the final analysis, yielding a general response rate of 97.4 per cent. Response rates did not vary between urban and rural areas.

It is important to highlight that out of the total 6,969 households interviewed in the first round (baseline), only 6,943 remained for the analysis. The reason for excluding 26 households is guided by problems (falsification) in the listing of households in two enumeration areas, namely, EA 19 (November) and EA 30 (December) that were detected in the midline (Round 2). Therefore, the households of each of both enumeration areas were excluded (24 in total) and the weights for that governorate were adjusted accordingly. In addition, 526 individuals were erroneously included as household members in the baseline according to follow-up information provided by the Round 2. When excluding them from the database, one household was lost. Finally, one additional household was excluded since only information from housing characteristics section was recorded.

Table M.1: Number of households and individual interviews, Yemen 2012

	Area of residence		Total
	Urban	Rural	
Number of households			
Original sample	1619	5533	7152
Final analysis sample	1568	5375	6943
Response rate (%)	96.8	97.1	97.1
Number of individuals			
Original sample	11,103	39,297	50,400
Final analysis sample	11,052	39,139	50,191
Response rate (%)	99.5	99.6	99.6

Source: NSPMS, Round 1.

Concluding remarks:

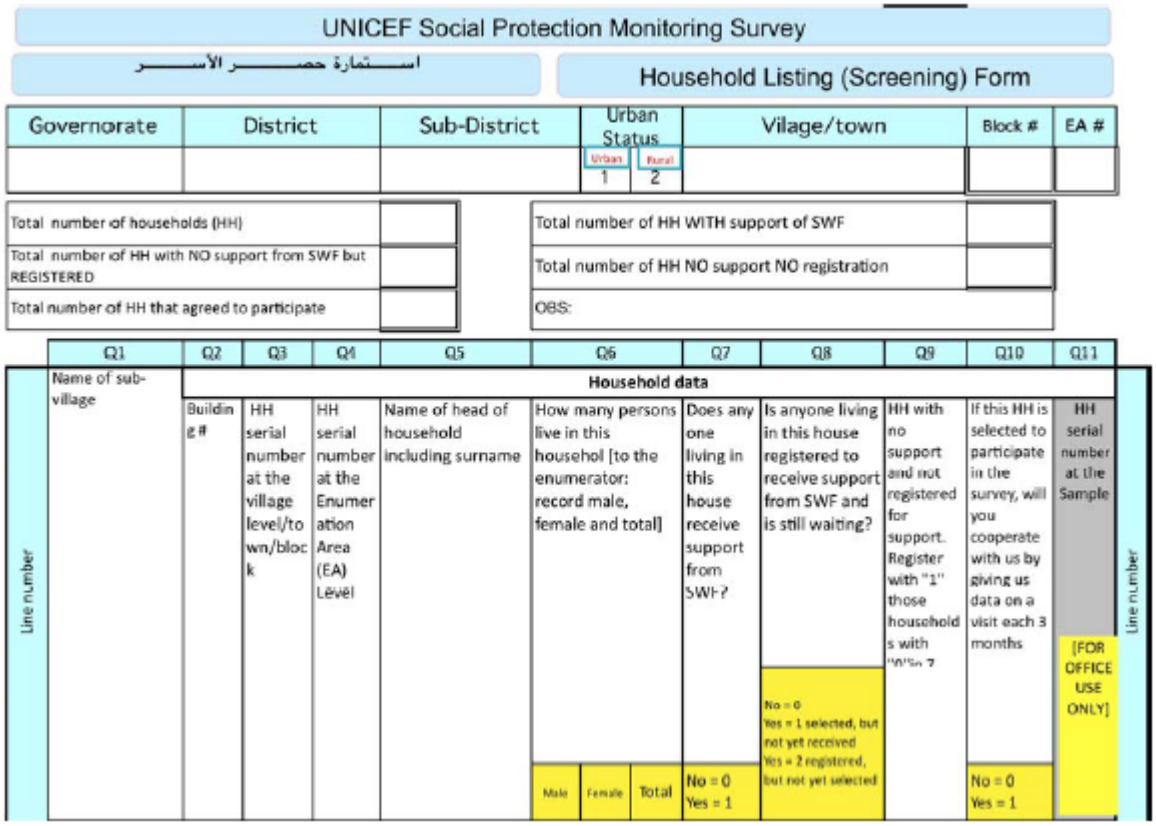
This section provided information on the process used to build the NSPMS sampling design, cross-sectional sampling weights, and the relevance of appropriately addressing item non-response problems. This information is useful for the release of official estimates based on the first-wave survey data. These estimates may be improved by incorporating available auxiliary information into the estimators' functional form – for instance, by further adjusting the cross-sectional sampling weights. The efficiency of these procedures depends on the quality of the relationship between the auxiliary and response variables; therefore, the analysts must choose the auxiliary information carefully.

The NSPMS will also be able to produce estimates for changes over time, based on longitudinal sampling weights. However, the calculation of these sampling weights depends on the information regarding the forthcoming Rounds of the survey; therefore, details about its building process will be documented in future report.

A final remark should be made on the absence of bias of the statistical estimation process for the NSPMS survey. Since a probability sample was selected, based on a measurable sampling design, the adoption of a Horvitz-Thompson estimator for totals, means and proportions ensures absence of bias of the estimates and their respective variances. Furthermore, under the assumption of a missing at random non-response mechanism, the corrections applied to the basic sampling weights also let the Horvitz-Thompson estimators based on the corrected weights to retain the same desired statistical properties (Isaki and Fuller, 1982; Vieira, 2009).

Annex:

A. Listing questionnaire used in the screening process



UNICEF Social Protection Monitoring Survey

استمارة حصر الأسر

Household Listing (Screening) Form

Governorate	District	Sub-District	Urban Status (Urban 1 / Rural 2)	Vilage/town	Block #	EA #

Total number of households (HH)		Total number of HH WITH support of SWF	
Total number of HH with NO support from SWF but REGISTERED		Total number of HH NO support NO registration	
Total number of HH that agreed to participate		OBS:	

	Q1	Q2	Q3	Q4	Q5	Q6			Q7	Q8	Q9	Q10	Q11	
	Name of sub-village	Household data												
Line number		Building #	HH serial number at the village level/town/block	HH serial number at the Enumeration Area (EA) Level	Name of head of household including surname	How many persons live in this household [to the enumerator: record male, female and total]			Does any one living in this house receive support from SWF?	Is anyone living in this house registered to receive support from SWF and is still waiting?	HH with no support and not registered for support. Register with "1" those households with "0"s ?	If this HH is selected to participate in the survey, will you cooperate with us by giving us data on a visit each 3 months	HH serial number at the Sample	Line number
						Male	Female	Total	No = 0 Yes = 1	No = 0 Yes = 1 selected, but not yet received Yes = 2 registered, but not yet selected		No = 0 Yes = 1	[FOR OFFICE USE ONLY]	

Map 1. Districts where at least one Enumeration Area (EA) has been selected for the sample