

IHSES SAMPLING DESIGN

Total sample size and stratification

The IHSES and the MICS share a common sample of 18,144 households: 324 households in each of 56 major strata. The strata are the rural, urban and metropolitan sections of each of Iraq's 18 governorates, with the exception of Baghdad, which has three metropolitan strata.

Sampling strategy and sampling stages

The sample was selected in two stages, with groups of majals (Census Enumeration Areas) as Primary Sampling Units (PSUs) and households as Secondary Sampling Units.

In the first stage, 54 PSUs were selected by systematic equal probability sampling (*seps*) within each stratum, using the number of households recorded by the 1997 Census as a measure of size.

In the second stage, a sample of six households was selected by systematic equal probability sampling (*seps*) within each PSU. To these effects, a cartographic updating and household listing operation was conducted in 2006 in all 3,024 PSUs, without resorting to the segmentation of any large PSUs.

The total sample is thus composed of 6 households in each of 3,024 PSUs.

Trios, teams and survey waves

The PSUs selected in each governorate (270 in Baghdad and 162 in each of the other governorates) were sorted into groups of three neighboring PSUs called *trios* – 90 trios in Baghdad and 54 per governorate elsewhere. The three PSUs in each trio do not necessarily belong to the same stratum.

The 12 months of the data collection period were divided into 18 periods of 20 or 21 days called survey waves. Fieldworkers were organized into teams of three interviewers, each team being responsible for interviewing one trio during a survey wave. The survey used 56 teams in total – 5 in Baghdad and 3 per governorate elsewhere. The 18 trios assigned to each team were allocated into survey waves at random.

Contingency procedures

The IHSES will not replace any of the randomly selected units (PSUs or households.)

However if a survey team is unable to visit a trio of PSUs in the originally allocated wave, that trio will be swapped with the trio from a randomly selected future wave that can be visited. This is an emergency procedure intended to be used when dictated by security reasons.

Near the end of the survey period, some of the teams may find that none of the remaining trios is secure enough to be visited in the present wave. In this case, they will randomly select instead one secure trio, and visit in each of its clusters a new *seps* sample of six households (different from those interviewed when the trio was originally visited in the past).

Selection probabilities and sampling weights

The selection probability p_{hij} of household hij in PSU hi of stratum h is given by

$$p_{hij} = \frac{k_h n_{hi}}{N_h} \times \frac{m_{hi}}{n'_{hi}} \dots\dots\dots(1)$$

where

k_h is the number of PSUs selected in stratum h (nominally always 54;)

n_{hi} is the number of households in PSU hi , as per the 1997 Census;

N_h is the total number of households in stratum h (also as per the 1997 Census;)

m_{hi} is the number of households selected in PSU hi (nominally always 6;) and

n'_{hi} is the number of households in PSU hi , as per the 2006 listing operation.

The two factors on the right-hand side of Equation (1) represent, respectively, the probability of selecting PSU hi in stratum h , and the conditional probability of selecting household hij in PSU hi .

To obtain correct estimates from the survey, the data from each household should be affected by a sampling weight (or raising factor) w_{hij} , equal to the inverse of its selection probability p_{hij} . if

- the number of PSUs visited in each stratum and the number of households visited in each PSU were always equal to their nominal values (54 and 6, respectively.) In other words, if the teams never had to resort to the trio-swapping tactic described above; and if
- the number n'_{hi} of households listed in 2006 were equal to the number n_{hi} of households recorded by the 1997 Census in all PSUs;

then Formula (1) would simplify to $p_{hij} = 324 / N_h$ – a constant for each stratum, and the sampling weights would become simply $w_{hij} = N_h / 324$. In other words, the sample would be self-weighted inside each stratum, with sampling weights equal to the inverse of the sampling fractions.

Nonresponse

The survey may not be able to collect data from all of the selected households. This can happen either

- [a] because whole PSUs cannot be visited, or
- [b] because in certain PSUs some households cannot be interviewed.

Evidence from almost a thousand PSUs visited during the first six waves shows that in the IHSES problem [b] will be minimal – the proportion of PSUs where 6 or 5 households were interviewed are, respectively, 92.4 and 6.7 percent – more than 99 percent in total. Some of the interviewed households may be “lost” at the analytic stage, but everything seems to indicate that the fieldwork quality control procedures put in place by the survey have been effective and these cases will be few. Under these conditions, the best assumption is the simplest one – consider that the households present in the final datasets in each PSU are an equal-probability sample of all

households in the PSU, and that factor m_{hi} in equation (1) represents the number of useful households in the PSU (rather than the number of households selected in the PSU.)

Problem [a] cannot be treated so lightly. It is possible that in spite of the trio-swapping tactic, some PSUs cannot be visited at all at the end of the fieldwork period, because of security reasons. Assuming that factor k_h in Equation (1) represents the number of PSUs actually visited by the survey in the stratum (rather than the number of PSUs selected in the stratum) would be wrong, because that would be assuming that insecure areas are similar to secure areas, which is obviously not the case. It is more appropriate to preserve the original meaning of factor k_h , and always use the nominal figure of 54 PSUs per stratum, even if less than 54 can be visited in the end. This is equivalent to acknowledge that the survey could only visit secure areas.

Analysis of preliminary datasets

Since trios were allocated into waves randomly, the sub-sample observed by the survey in any sub-annual period Δt could be considered in principle as an equal-probability sample of the total, and analyzed with weights obtained by simply multiplying the weights from Equation (1) by the adjustment factor $18/\omega$, where ω is the number of waves in Δt .

However, since the survey was not fielded in some of the governorates during some of the waves, and because the incidence of insecurity will not be known with certainty until the end of the survey, it is preferable to use instead stratum-wise adjustment factors computed as the ratios between the current total number of households in the stratum and the sum of the weights of all households from the stratum in the preliminary dataset. The best estimate for the current total number of households in each stratum can be obtained from the survey's own 2006 cartographic updating and household listing operation, as $N'_h = (N_h/k_h) \sum_i (n'_{hi}/n_{hi})$, where the terms on the right hand side have the same meaning as in Equation (1).