1 Sample design and estimation

The sample of villages for CSES 2011 is just a simple random 50 % subsample from the CSES 2009 sample of villages. Consequently, the description of the CSES 2011 sample design will by necessity begin with a description of the CSES 2009 design.

The sample is selected in three stages. In stage one a sample of villages is selected, in stage two an Enumeration Area (EA) is selected from each village selected in stage one, and in stage three a sample of households is selected from each EA selected in stage two.

Different aspects of the CSES 2009 sample design are described in the following sections. The CSES 2011 subsample and the method of calculating sampling weights is described in sections 1.5 and 1.6.

1.1 Target population, sample frame of villages

The target population for CSES is all "normal" households in Cambodia. The term normal is defined in the Population Census 2008 as households that are not institutional households, homeless households, boat population households or households of transient population. (Institutional households are boarding houses, military barracks, prisons, student dormitories, etc.). Preliminary data from the General Population Census 2008 was used to construct the CSES 2009 sampling frame for the first stage sampling, i.e. sampling of villages. All villages except 'special settlements' were included in the frame. In all, the first stage sampling frame of villages consisted of 14,073 villages.

1.2 Stratification, allocation of the sample over strata

The sampling frame of villages was stratified by province and urban and rural. In total there are 48 strata. Each stratum of villages was sorted by district, commune and village code.

For the CSES 2009 survey it was decided to have a sample of 720 villages. The total sample size was divided (stratified) into to two: one sample size for urban villages and the other for rural villages. The calculation of the sample sizes for urban and rural areas were done using the proportion of consumption in the two parts of the population. Data on consumption from the CSES 2007 survey was used. The resulting sample sizes for urban villages was 240 and for rural 480.

The allocation of urban and rural sample size over provinces was done so that each province got its proportional share (approximately) of the sample.

1.3 Allocation of the sample over survey months

The total sample of 720 villages was divided into 12 monthly samples of equal sizes. The monthly samples consisted of 20 urban and 40 rural villages. The division of the annual sample into monthly samples was done so that as far as possible each province would be represented in each monthly sample. Since the sample size of villages in some provinces is smaller than 12, all provinces were not included in all monthly samples. Also, the outline of the fieldwork with teams of 4 enumerators and one supervisor puts constraints on how to divide the annual sample into monthly samples. The supervisors must travel between the villages in a team and therefore the geographical distance between the villages surveyed by a team cannot be too large.

1.4 Sampling

The sample was selected in three stages:

Stage 1. A random sample of villages was selected from each stratum. The sampling method can be expressed in technical terms as: "*without replacement systematic sampling with probabilities proportional to size*". The size measure used was the number of households in the village according to the sampling frame. The selection of villages was done at NIS.

Stage 2. One EA was selected by Simple Random Sampling (SRS), in each village selected in stage 1.In a few large villages more than one EA was selected. The selection was done at NIS.

Stage 3. In each selected EA a sample of 10 households (urban villages) or 20 households (rural villages) was selected.. The selection of households was done in the field. All households in selected

EAs were listed by the enumerator. The sample of households was then selected from the list by systematic sampling with a random start (the start value controlled by NIS).

The sampling resulted in a sample of 12,000 households, 2,400 urban households and 9,600 rural households.

1.5 Sample design and sampling for CSES 2011

The sample design for CSES 2011 is basically the same as the CSES 2009 design. For the 2011 survey a subsample of 360 EAs (stage 2 units) was selected from the CSES 2009 sample of 720 EAs. The selection was done by simple random sampling within strata. The selection resulted in 136 urban EAs and 224 rural EAs. It is the same EA:s as in 2010.

Households were selected in the same way as in CSES 2009. For CSES 2010 and 2011 only 10 households are selected in each rural EA.

The sampling resulted in a sample of 3,600 households, 1,360 urban households and 2,240 rural households.

1.6 Sampling weights for CSES 2011

The 3,600 households in the sample did not have the same probability of being selected to the sample. Urban households had on average a 1 in 400 chance of being selected while rural households only had a 1 in 1000 chance of being selected. Urban households are over-represented in the sample as a result of this way of selection. This is not a flaw in the design but rather an intended feature.

The over-representation of urban households in the sample must be compensated for in the calculations of results from the sample. Each household must be assigned a "sampling weight" that reflects the chance (probability) of the household to be selected to the sample.

The sampling weights were calculated in two steps:

Step 1, Preliminary weights: The probability of being selected to the sample was calculated for each household, giving the preliminary sampling weight as the ratio *1/probability* (=inverse of the probability).

Step 2, Final weights: The preliminary sampling weights were added over all sample households within each stratum. The sum of the weights is an estimate of the total number of households in the stratum. This estimate was compared to the number of households according to demographic projections based on the 2008 Population Census. The preliminary sampling weights were then "calibrated" so that the sum of the weights should agree with the demographic projections.

2. Quality of the estimates from CSES

All survey data are subject to errors from various sources. The errors may occur at any stage during the survey work. A broad fundamental distinction of errors is between sampling errors and non-sampling errors. The quality of an *estimate*, i.e. a result, from the survey is a function of both sampling and non-sampling errors.

2.1 Sampling errors

There is always an uncertainty in the results (estimates) from the survey due to the fact that not all households in Cambodia are included in the survey. This uncertainty is indicated by the standard error for the estimate. A large standard error implies a large uncertainty in the estimate. The uncertainty can also be expressed as a *confidence interval* ("margin of error") around the estimate. The confidence interval around the estimate is the interval obtained by subtracting two standard errors from the

estimate (=lower boundary of the interval) and adding two standard errors to the estimate (=upper boundary of the interval)[1]. The confidence interval is an interval within which the true value for the population can reasonably be assumed to be. An example:

> The estimated average floor area of residential houses/dwellings for the households in Cambodia is 44.5 square meters (sqm). The standard error is 0.77 sqm. The confidence interval becomes 44.5 +/- 2*0.77 which results in the interval [43.0 - 46.0]. This interval covers the true, unknown, average floor area for all households in Cambodia with a high degree of confidence.

Standard errors or confidence intervals are presented for some important estimates in the report. Furthermore, in some of the diagrams the confidence intervals are superimposed. The standard errors have been calculated by the Taylor linearization method. The software used was Stata 11, survey data analysis (svy) module.

2.2 Non-sampling errors

Non-sampling errors are mainly associated with field work and data processing procedures. The non-sampling errors in CSES are non-response errors, response errors and data processing errors. Table y gives an overview of the different types of error and presents an assessment of the effects of the errors on survey results.

Type of error	Description	Assessment
Non-response errors	Some of the selected households do not participate in the survey because they refuse or are not available for interview. Also *partial nonresponse* where the household cannot or does not want to answer a question	The non-response rate is very low; only eight households out of the selected 3,600 households are missing from the survey. Therefore, the effects of non-response errors is negligible in CSES 2011
Response errors (measurement errors)	Errors in responses from the households because the household: - doesn´t understand the question correctly - doesn´t know the correct answer, or doesn´t remember correctly - doesn´t want to give the correct answer (on sensitive questions) - gets tired of the questions and doesn´t want to cooperate fully during the whole interview. Errors can also be caused by the interviewer when he/she doesn´t record the responses correctly	It is very difficult to assess the response errors that arise in the survey. Some response errors are found and corrected in the automatic logical checks and range checks that are done at data entry and right after data entry. Some other errors present in the survey cannot be detected unless special quality studies are carried out (re-interview studies, register studies, "data confrontation"). This has not been done. The CSES has been carried out four times prior to the present survey. Over the years errors and ambiguities in questions, definitions and concepts have been addressed and corrected. It is therefore fair to say that many sources for potential response errors have been eliminated. Still, there are errors left in the data. These errors have limited impact on most estimates but may have rather large impact on some estimates, for example estimates of expenditure on commodities with low-frequent purchases.
Data processing errors	Data entry staff make mistakes; the staff coding the answers to the open-ended questions (like *occupation*) put wrong codes in some cases	A large number of automatic logical checks and range checks are done at data entry and right after data entry. Also, the staffs analyzing the data carry out additional checks of outlier values and other values that are clearly inconsistent. The thorough editing of the data makes sure that most of the substantial data processing errors are detected

[1] The theoretically correct method is to add and subtract 1.96 standard errors

		and corrected – except for the coding errors. The coding errors can only be detected by special studies like re-coding by another coder and reconciliation of differing codes. No such study has been made but great efforts have been made to train the coders properly. This has for sure reduced the level of coding errors considerably.