

*Vietnam Household Living Standards Survey (VHLSS), 2002 and 2004<sup>1</sup>*

*Basic Information*

---

<sup>1</sup> *This document was prepared by Phung Duc Tung (pdtung@gso.gov.vn) & Nguyen Phong (nphong@gso.gov.vn) from Social & Environment Statistics Department- Vietnam General Statistics Office.*

## CONTENTS

1	Overview .....	1
2	Survey Questionnaires.....	2
2.1.	VHLSS 2002 questionnaires .....	2
2.1.1	Household Questionnaire .....	2
2.1.2	Commune/Ward Questionnaire .....	7
2.2.	VHLSS 2004 questionnaires .....	9
2.2.1	Household questionnaire .....	9
2.2.2	Commune questionnaire.....	14
3	Sample.....	15
3.1	Sample design.....	15
3.1.1	Target population.....	16
3.1.2	Three-stage design.....	16
3.1.3	Sample Size .....	18
3.1.4	Cluster size .....	21
3.1.5	Stratification .....	23
3.1.6	Eligibility and response rates.....	24
3.1.7	Sample rotation.....	24
3.2	Sampling procedures .....	24
3.2.1	Selection of the PSU.....	24
3.2.2	Selection of secondary sampling units .....	25
3.2.3	Selection of households.....	25
3.2.4	Sample selection for the augmented sample for 2002.....	26
3.3	Revisions to sample design for the 2004 VHLSS .....	26
3.3.1	Expected sampling errors under different assumptions on sample sizes.....	26
3.3.2	Expected sampling errors at provincial level for different sample sizes .....	28
3.3.3	Changes to sample size for the 2004 survey.....	30
3.3.4	Differences between the 2002 and 2004 sample design.....	31
3.4	Sampling evaluation .....	32
3.4.1.	Change in definition of households.....	32
3.4.2	Comparability of sample population estimates to population projections.....	32
3.4.3	Design effects and rates of homogeneity.....	32
3.5	Weights and standard errors for the VHLSS sample.....	34
3.5.1	Calculation of sampling weights .....	34
3.5.2	Calculation of standard error .....	36
4	Implementation.....	39
4.1.	Central level .....	39
4.2.	Provincial level.....	39
4.3.	District level .....	39
4.4.	Commune/ Ward level.....	39
4.5.	Recruitment interviewers and team leaders.....	39
4.6.	Training of the trainers .....	39
4.7	Training interviewers and team leaders.....	40
4.8.	Media campaign .....	40
4.9.	Field work.....	40

5	Data processing .....	41
6	Dissemination policy .....	41
6.1.	Principles in VHLSS Data Dissemination.....	41
6.1.1.	Legality of data dissemination.....	41
6.1.2.	Equitability in data dissemination .....	41
6.1.3.	Maintenance of confidentiality .....	42
6.1.4.	Guarantee of data quality.....	42
6.2.	The Policy of VHLSS Dissemination.....	42
6.2.1.	Data to Be Disseminated .....	43
6.2.2.	Recipients and Dissemination Mechanism.....	45

## **Vietnam Household Living Standards Survey (VHLSS) 2002 and 2004**

### **Basic Information**

#### **1 Overview**

During the 1990s two important household surveys were developed in Vietnam to look at living standards: the Multi-purpose Household Survey (MPHS) and the Vietnam Living Standards Survey (VLSS). The MPHS is a key component of the overall statistical system and survey program as information collected is used for multiple purposes including CPI calculations, national account statistics and poverty estimates among others.

The MPHS survey has been implemented every 1 to 2 years since 1994 and had a sample size from 25,000 to 47,000 households. The focus of the survey has been household income and expenditures; however other important aspects of living standards such as education, health or employment have been included in various rounds. The large sample size, limited budget, large number of personnel involved, short time for implementation, unclear medium and long-term plans, significant changes in methodology and content over time and lack of standardization in methodology have led to some inconsistencies in data collection over time, inability to make clear analysis plans, and possibly have had negative impacts on data quality. In addition, the MPHS has not been disseminated in a timely manner nor analyzed in depth.

The two VLSS were implemented in 1992-93 and 1997-98 with funding from the UNDP and Sida and with technical assistance from the World Bank. They provided important opportunities for the General Statistics Office (GSO) to learn a new survey methodology of integrating many topics into one survey as well as exposure to up-to-date methods of questionnaire design and quality control. It also provided a high quality data set covering a great variety of topics that has been widely used in social policy research. However, with the given budget and long, detailed questionnaire only a relatively small sample (4800 households in 1992- 93 and 6000 households in 1997- 98) could be selected which was not large enough to provide information at a level that satisfied the government or provinces. In addition, the frequency of implementation every five years was too low to satisfy the need for up-to-date information to monitor social issues.

The implementation of these two different surveys with similar purposes has led to duplication and waste. It is therefore a high priority of the GSO to integrate the MPHS and the VLSS surveys and create a medium term plan for implementing a new VHLSS through 2010. With the technical assistance from UNDP and World Bank, GSO has developed a strategy for the VHLSS over the next 10 years (2000- 2010) that transformed the VHLSS into a biennial (once every two years) core and rotating module household survey. The VHLSS strategy is approved and it is implementing as follow:

<i>Year</i>	<i>Survey content</i>
2002	Expenditure and income (core) + basic information of other sections
2004	Core + Land and Non farm activities (rotating module)
2006	Core + Health and Education (rotating module)
2008	Core + Land and Non- farm activities
2010	Core + Health and Education

This document provides background information on the Vietnam Household Living Standards Survey 2002 and 2004, and the data-sets resulting from this nation-wide household survey. Information is provided on the survey design and implementation, the content of the different questionnaires, data processing activities and data dissemination policy for this survey.

## **2 Survey Questionnaires**

### **2.1. VHLSS 2002 questionnaires**

The VHLSS 2002 used 3 questionnaires: short household questionnaire (excluded most of consumption expenditure information), long household questionnaire (including detail consumption expenditure information), and commune questionnaire.

#### ***2.1.1 Household Questionnaire***

The short household questionnaire contains 9 sections each of which covered a separate aspect of household activity.

##### **COVER PAGE (SURVEY INFORMATION)**

- 1. HOUSEHOLD ROSTER**
- 2. EDUCATION**
- 3. EMPLOYMENT**
- 4. HEALTH**
- 5. INCOME AND HOUSEHOLD PRODUCTION**
- 6. EXPENDITURE (collected only for long questionnaire)**
- 7. DURABLE GOOD AND ASSET**
- 8. HOUSING**
- 9. PARTICIPATION IN POVERTY REDUCTION PROGRAMS**

The individual designated by the household members as the household head provided responses. For some others (education and health expenditure, income from wage, agro-pastoral activities, non-farm self-employment, food expenditures, non-food expenditures) a member identified as most knowledgeable in each those sections provided responses.

Unlike the VLSS, the household questionnaire of VHLSS 2002 was completed in one interview. The survey was designed so that more sensitive issues such as credit, savings and assets were discussed near the end. The content of each module is briefly described below.

##### **Section 0 COVER PAGE**

The cover page collected information about the household identification codes (province, district, commune, enumeration area and household codes), the religion, and ethnic group of the household, language used by the respondent and codes of interviewer and team leader as well as the date of interview.

##### **Section 1 HOUSEHOLD ROSTER**

The roster in section lists the sex, relation to head of household, birth dates, age and marital status (for people aged 13 and over) of all people who are living in the household since at least 6 months ago. The household head is listed first and receives the personal id code 1. Household members were defined generally to include “all people who normally live and eat their meals together in this house and have done so for 6 or more months out of the past year”. However, there are 7 exceptions:

The head of household is always considered as household member even he or she does not live and eat in the household more than 6 months

1. Infant less than 6 months old is still considered as household member
2. The people who are going to live in the household for a long time such as daughter-in-law, son-in-law, relatives who came back from outside (retired, soldier ect...) are considered as household member even they are living less than 6 months  
Students living outside the household but still supported by their family are considered as household member
3. Guests or relatives living with the household 6 or more months and eat their meals together are considered as household member
4. Hired workers, servants, or lodgers or guests if they are members of households elsewhere are not the household members
5. Individuals who died during the past 12 months or moved out of the household and do not intend to come back are not considered as household member

A transforming table between lunar and solar year is provided in the questionnaire to help respondents recall the year they were born.

### Section 2 EDUCATION

Section 2 first asks which grade an individual has completed in the school, and if yes it is less than grade 5 then ask whether she or he can read and write, the highest obtainable diploma and they are currently in school or have completed their schooling. Depending on the answer, the interview continues to ask details about school expenditures including tuition, contribution (construction fund, parent's association), uniforms, textbooks, other education equipments, extra learning (including foreign language and computer), other expenses and a total. Individuals were asked whether or not they were allowed exemptions or reductions in fees, and for what reason as well as the percentage of exemption for tuition and contributions. Students were also asked about the value of scholarships. The last question in this section is to ask individuals about the expenditure in the past 12 months on other courses outside of school such as study of foreign languages or computers.

### Section 3 EMPLOYMENT

All individuals age 10 and older were asked to respond to the economic activity questions in Section 3, beginning with questions on the nature of their work in the last 12 months. First question asks whether an individual has been worked as wage earner or self employment in agriculture or self-employment in non-farm activities. If they have worked in the past 12 months, the next question were collected on working hours in the past 7 days and then the main job in the past 12 months was collected information about occupation and industry of employment and type of employer, number of working months in the past 12 months, average number of working days per months in these months and average number of working hours per day. Then the same information about total number of working months, days per month and hours per days of total other earned jobs was collected. Finally, hours spent doing household chores per day and numbers of days working for community were collected for each household member age 10 and older. Occupation and industry of employment codes are printed directly in the household questionnaire.

#### Section 4 HEALTH

The section begins by asking respondents whether or not any individual in their family went to any health provider in the past 12 months (including health provider came to their home, health examination, vaccination with payment, pregnant examination, child delivery), and if yes, the next questions were asked about the name of the health providers, inpatient or outpatient and the total cost of each time for each inpatient or outpatient treatment.

The rest of questions were asked for all households about other health expenditures on medicines and health equipments, contribution for health fund for local community and expenditure on health insurance in the past 12 months.

#### Section 5 INCOME AND HOUSEHOLD PRODUCTION

5A INCOME FROM WAGE

5B FARM ACTIVITIES

5B1 AGRICULTURE, FORESTRY AND WATER- SURFACE LANDS

5B2 CROPS

5B2.1 FOOD CROPS

5B2.2 ANNUAL AND PERENNIAL INDUSTRIAL CROPS

5B2.2 FRUIT CROPS

5B2.4 CROP BY- PRODUCTS

5B2.5 CROP EXPENDITURES

5B3 LIVESTOCK

5B3.1 TURNOVER FROM LIVESTOCK

5B3.2 LIVESTOCK EXPENDITURE

5B4 AGRICULTURAL SERVICES

5B4.1 OUTPUT FROM AGRICULTURAL SERVICES

5B4.2 EXPENDITURE FOR AGRICULTURAL SERVICE ACTIVITIES

5B5 FORESTRY ACTIVITIES

5B5.1 OUTPUT FROM FORESTRY ACTIVITIES

5B5.2 EXPENDITURE ON FORESTRY ACTIVITIES

5B6 FISHERY ACTIVITIES

5B6.1 OUTPUT FROM FISHERY ACTIVITIES

5B6.2 EXPENDITURE ON FISHERY ACTIVITIES

5C. NON- FARM ACTIVITIES

5C1. OUTPUT FROM NON- FARM ACTIVITIES

5C2. EXPENDITURE ON NON- FARM ACTIVITIES

5D. OTHER INCOME SOURCES

5D1. OTHER INCOME

5D2. NON- INCOME SOURCES

In this section, most questions refer to the past twelve months. This section is by far the longest section of the household questionnaire, with many subsections that contain information on different aspects of income sources from wage, agricultural production and related livestock and fishery and non- farm activities.

Section 5A collects information of the household member who has been worked as employee in the past 12 months, starting with the questions on occupation, industry of employment and type of employer, total salary received and other incomes (received during Tet holiday, social subsidy for working accident, during pregnancy time ect, accommodation for going to business and other) during the past 12 months of the main job. The last two questions in this section is to ask whether the individual have been working as employee for second or more jobs and if yes, total amount he received during the past 12 months.

Section 5B1 collects information on household's control over different plots of land of different tenures. These include their own land, auction land, rented land, and land rented out by the household. Annual crop and perennial crop land, forestry land, water surface and unused land are included. In each case data are obtained on total area, area for long use purpose, area having land use certificate (LUC) and who has the name on the LUC.

Section 5B2.1 to 5B2.3 contains detailed output information for all crops grown by the household. This information is obtained separately for each crop and includes (in most cases) information on total area planted, quantity harvested, amount and value sold out and the main buyer, and total value of output in the past 12 months. Section 5B2.4 contains information about total value of crop BY-products during the past 12 months, including straw, thatch, sweet potato leaves and sterms, sterms of cassava plant ect.

Section 5B2.5 contains detailed information on use of production inputs including seeds, manure, chemical fertilizer, insecticides and herbicides, small equipment, fuel including electricity, gasoline, repair and maintenance, depreciation of equipment and assets, expenditure on rented land and equipments, labor, irrigation fee, pay the interest for loan, agricultural tax and other expenditures (fee, post, insurance, ect) for the 4 main groups of all crops cultivated by the household. This information is crop-specific for each main group of crops so we could link with output information in the earlier sections.

Section 5B3.1 collects information on the output of livestock, poultry and other animals that are sold out, consumed by a household, used for other purposes and total amount and value of output in the past 12 months.

Section 5B3.2 collects information on the cost of livestock, poultry and other animals specified by animal feed, equipment, fuel, repair and maintenance, depreciation of assets, renting land, labor, medicine, pay interest, tax and other expenditure.

Section 5B4.1 and 5B4.2 collects information on revenue and expenditure of agricultural service activities of the household. These activities include Plowing/ land preparation service, irrigation, primary crop processing service and other service.

Section 5B5.1 and 5B5.2 collects information about output and cost of forestry activity of the household in the past 12 months for 10 main type of tree including mu oil tree, cinnamon tree, anise tree, pine tree, varnish tree, tree for wood, bamboo, fan falm tree, water coconut falm, other forestry tree and wood. The cost items in section 5B5.2 are similar to the section 5B2.1

Section 5B6.1 and section 5B6.2. collect information about output and cost of aquaculture production in the past 12 months including quantity and value, quantity and value that household consumed and sold out, cost of raising, catching and service for breed, foods, non- durable items, energy fuel, small repair and ect..

Section 5C1 collects information about non- farm activities of the household in the past 12 months, including type of activity, industry code, the main person who operated, turnover, values that household consumed in the past 12 months, and the cost for each activity including material, electricity, small repair and maintenance, depreciation of assets, labor cost, interest, tax and ect...

Section 5D1 collects information about income from other sources in the past 12 months. The other household income includes remittance and value of in- kind received from both domestic and oversea, Pension, sickness and one-time job loss allowance, social insurance allowance, other income from social insurance, interest of saving, dividend, income from renting out workshop, machinery, assets, income from wining lottery, charity and support from other organization.



Section 5D2 collects information about selling machine, equipment, workshop, house, gold, and withdrawal from saving, stock, ect...

#### Section 6. EXPENDITURE

Section 6 part A collects information about expenditure of the household on food and drink during holidays (section 6 part A1) and daily (section 6 part A2). Section 6 part A1 collects information about quantity and value of 24 main food items that Vietnamese households usually consume during the holidays, especially in Tet holiday (new lunar year). Those food items are mainly typical meats such as pork, beef, chicken; special and glutinous rice; typical drink such as coffee, tea, beer, wine and alcohol, fruit, outside eating that household brought or self made or received during the holidays in the past 12 months. Section 6 part A2 collects information about the household expenditure on the daily food and drink of 58 food items, excluding the amount that household consumed during holidays. It collects detailed information on market purchases and consumption from home production. Information is obtained on the number of months (in the past 12 months) each food item was purchased, the number of times purchases were made during those months, the quantity purchased each time, and the value per purchase. These four pieces of information can be combined to obtain the total expenditure on food in the 12 months before the data of the interview. Besides market purchases (including barter), information is also collected on consumption from home production. Again data are obtained on the number of months each item was consumed, but unlike market purchases, the information of the quantity and value of consumption is obtained by asking a single question on the total amount for the last 12 months (as opposed to asking how often purchased each month, quantity purchased each time).

Section 6 part B collects information on non-food household expenditures. In section 6 part B1 respondents were asked to recall number of months purchased, number of time purchases made during those time and total amount spent in the past 12 months as well as the value of self making on daily expenses such as lottery tickets, cigarettes, soap, personal care products, cooking fuel, matches and candles, and gasoline, ect... In section 6 part B2 collects information about annual expenditure for 32 items such as fabric, ready made clothing, mosquito net, face towel, scarves, rush mats, blankets, pillows, tailoring or laundry service, shoes, nylon sheeting, light buds, electric wire, ect. The data are obtained on the value purchased and self- making in the past 12 months. Section 6 part B3 collects information about amount of money that household spent during in the past 12 months for contributions to various funds, public labor contribution, all kinds of taxes (excluding production taxes), wedding, funeral on special occasions of household members, parties (celebrate birthday), give, donate, support (cash and kind) and others. Section 6 part B4 collects information on other household expenses that are not considered as household expenditure in the past 12 months, including lending money, pay debt, return advance (including payment of interest), contribute to revolving credit group, buying share, certificate and stock, purchase gold, silver, precious stone, foreign currency for saving purpose, saving account, life and security insurance, outstanding investment, other expenditure.

#### Section 7. FIXED ASSETS AND DURABLE GOOD

This section collects information about 59 kinds of fixed assets and durable goods of the households, including perennial crops garden, aquaculture production area, other production land area, buffalo, cow, horse for production and breeding, feed grinding machine, rice milling machine, car, trailer, motorbike, wagon, boat with engine, computer, Television, telephone, air-conditioner, ect... Each asset or good was asked about the month and year household received, value at the time household brought or received, the value at the interview time, the using purpose, the share (percentage) using for production purpose and the share using for consumption purpose.

### Section 8. HOUSING

This section collects information about living condition of the household that includes the number of houses or flats that household is living, total living area, type of the main dwelling (quality of the house), the time of living in those houses or flats, the place that household has been living before the current place, the owning status of the current living place, amount of money paid during the past 12 months for renting of the houses or flat, the value of the house or flat at the interview time. In addition, this section also collects information about the other house or flat that household owned and whether the household received money from renting it out, amount received from renting in the past 12 months and the value of that house or flat at the interview time. The household was also asked to provide information about expenditure on construction, repair, renovation and improvement of the house during the past 12 months, the main drinking source and type of toilet that the household is currently using, amount paid for using the drinking water and the main source of lighting and the method that household uses to dispose the garbage in the past 12 months. Furthermore, this section collects information about the access of household to Vietnamese television, radio channels and the popular newspapers in Vietnam as well as access to internet in the past 12 months.

### Section 9. PARTICIPATION IN POVERTY REDUCTION PROGRAMS

This section collects information with the purpose to examine the coverage of National Target Program to the poor households and their accession to the credit, health programs. In the first three questions, the household was asked in order to define whether the household belongs to family of invalids, sick war veterans, martyr, Vietnamese heroic mother or lonely elderly, disabled or poor household and then to examine whether the household received the Poor Household Certificate or not. In addition, the household received the Poor Household Certificate was asked to define whether they have got the exemption from health consultation and treatment or received the assistance from the State mass organizations, associations for repairing or constructing dwelling, or they have got exemption from agriculture land use tax in year 2001. The information about the loans, value of the loans, the interest rate from different banks and social funds, private borrowers are collected in the last 4 questions of this section.

#### *2.1.2 Commune/Ward Questionnaire*

The commune questionnaire includes 9 sections and was administered by the team leader and completed with the help of village chiefs, teachers, government officials and health care workers. The questionnaire was administered in both rural and urban areas but some section was only collected in rural area such as non- farm employment opportunities and infrastructure and transportation. The commune questionnaire contains 9 sections including:

0. SURVEY INFORMATION
1. MAIN CHARACTERISTICS OF THE COMMUNE/ WARD
2. GENERAL ECONOMIC CONDITIONS AND AID PROGRAMS
3. NON-FARM EMPLOYMENT OPPORTUNITIES
4. AGRICULTURE
5. PHYSICAL INFRASTRUCTURE AND TRANSPORTATION
6. EDUCATION
7. HEALTH
8. PUBLIC DISORDER AND OTHER SOCIAL AFFAIRS

**Section 0** contains basic background information on respondents including position in the commune, tenure in the current position, age, education, gender and ethnicity.

**Section 1** collects information on population and geographic region of the commune. It also collects information on land area, number of households, and number of agricultural households, the main ethnicities and religions of the population whether the commune was listed in the 135 program or considered as remote area by regulation of the government.

**Section 2** contains information on major sources of income in the commune, the changes of living standard and reasons for changes in standard of living of the people of the commune, types of aid programs received in the past 12 months and from whom, types of ad hoc aid, the project or programs of the government or other organizations are currently implementing in the commune such as job creation, poverty alleviation, infrastructure ect, number of households considered as poor by MOLISA method, number of poor households receiving various types of assistance such as subsidized credit, exemption or reduction school fee, relief fund for natural disasters and pre-harvested shortage, occupational and agricultural technology training, exemption from production tax.

**Section 3** collects information on the existence and numbers of enterprises/factories, cottage industries or handicrafts operating within 10 km from the commune center, types of enterprises, number of enterprise having more than 20 employee, numbers of enterprise located in the commune, the traditional businesses existing in the commune, the ownership of these traditional businesses, the time that enterprise was set up, number of employee, the average wage for these employee..

**Section 4** asks questions about the land distribution of the commune, main crops grown in the area, number of crops harvested each year and whether the productivity of the crops was increased or decreased during the past 12 months and the reasons for that. This section also collects information about types of land of the commune, total area for each type of land, percentage of irrigation and quality for annual planted land area and aquaculture water surface land. Daily wage rates for various agricultural works for men, women and children under 15 were collected. The existence and distance of agricultural extension center, and number of visits of extension workers were also asked about. This section also asked question about the main problems faced by farmers.

**Section 5** collects information at the village level about roads and waterways, the distance from village to the nearest road for car, number of months per year that road was not used by car, type of road surface of the road coming to the village, and the same information for waterway, and the distance and time from the surveyed village to various places such as the commune people's committee building, district center, post-office, telephone, food store and market. Questions were also asked about the main source of drinking water for most of people in the commune in the dry and rainy reasons, the availability of electricity in the commune and frequency of electricity outages.

**Section 6** collected information on distance, method of transportation and time to the nearest schools of each level from surveyed village. For each level of schooling, the reasons why children dropped out and the main education obstacles of the commune were asked about. Information on anti-literacy programs, the time it started and number of participants by gender was also collected. For surveyed village, information on preschool was collected.

**Section 7** gathered information on the main illness/ diseases of concern this the commune, the using of health services, major problems with the health services in the commune and distance and time from the commune to each of various health facilities.

**Section 8** gathered information on the most important social issues in the commune such as drugs addiction, prostitution, gambling, drunk, theft, superstition, ect, the number of people in the commune related to these issues, of which the number of children under 16 years old, the number of people receives regular social assistance from commune.

## 2.2. VHLSS 2004 questionnaires

VHLSS 2004 questionnaires are developed based on the VHLSS 2002 questionnaires to ensure the comparability between two surveys. There are some changes in households and commune questionnaires that will be discussed in the following sections. The major changes in the VHLSS 2004 household questionnaire are two additional new modules for long household questionnaire that are found in section 9 (additional section) and 10 (non-farm self employment activities). This part of the report will mainly discuss the changes in the household and commune questionnaires in 2004 compared to questionnaires in 2002.

### 2.2.1 Household questionnaire

#### Section 0 COVER PAGE

The main difference in the cover page of household questionnaire in 2004 compared to it in 2002 is the information to define whether the household is surveyed in 2002 or not that includes the code of province, district, commune/ ward, enumeration area, household, quarter of interview in 2002 that could help us to define the panel households (further information provided at sampling section).

#### Section 1 HOUSEHOLD ROSTER

There are some changes in this section. The section is divided into two parts. The first part (part A) has some changes. The question asking about relationship with household head combines all kids of the household head into one code, and natural brothers and sisters with other relation into another code. The date of birth for each member in the household excluded the information about the day of birth. In this section, there are three additional questions to collect information about number of months that the household member actually live in the household in the past 12 months, the registration status of each household member and number of year and month that household member has live in the current city/ province. The second part (part B) is a new part to collect information about household member of the household that surveyed in 2002, including individual code, gender, age in 2002, currently living in the household or not, the individual code in 2004, the reason for moving out of the household. There is no change in definition of household member between two surveys.

#### Section 2 EDUCATION

There are a lot of changes in this section compared to 2002. The question asking about the highest diploma that household member received is coding in more detail and divided into two columns (one for education and other for professional training) for interviewer to fill in the questionnaire. There is one more answer option (school does not ask fee) in the question asking about the reason for school fee and contribution exemption and this question is divided into two columns (one for school fee and one for contribution). There are three additional questions to collect information about whether the household member is currently in school or not, the grade and types of school

(public or private) of the household member who is currently in school, the amount received from social organization.

### Section 3 HEALTH

In 2002, section 3 is employment section but it is moved into section 4 in 2004. Section 3 collects additional information compared to 2002 about whether any household member was sick or ill during the past 4 weeks before asking about the health status of her or him in the past 12 months, number of days in bed with the help of other people, number of days could not work as normal, whether she or he has health insurance and receives from whom, which health facility she or he uses and the reason to use that health facility. The questions asking about the cost of using health services are collected in more detail for each time of visit for inpatient and outpatient treatment. In addition, this section in 2004 also collect information about whether household has enough money to pay for health costs or not and if not, how they resolve this problem.

### Section 4 INCOME

This section of VHLSS 2004 combines section 3 and section 5 in VHLSS 2002. The part 4A includes information of section 3 and section 5A of VHLSS 2002.

In section 4 part A, the question “did you look for a job or would like to work in the past 7 days” is dropped out. However, this section collects in more detail information about the main job in the past 12 months and the second job in the past 12 months including the number of working months, number of working days per month and the number of working hours per day, number of working year. There are also additional questions to collect information about whether household member is the government official or not, the total amount received from the third job, total hours of housework in the past 12 months.

Section 5B1 in VHLSS 2002 is moved to the section 9 of VHLSS 2004. There are major changes in questions about agricultural production (section 4B) in VHLSS 2004. It is divided rice production out of other crop production. In addition, the rice crops are asked in more detail than in 2002. It collects more information about the total lost in the past 12 months, amount used as seed for next crop, amount used for raising livestock, amount used as gift and the rest. There is additional question to collect information about amount that household uses for consumption in other annual crops, annual and perennial industry crops, and fruit but it does not collect information about who the household sold these products out in 2004. Instead of collecting total value for each by-products of crops as in 2002, it collects information about the value of sold out, value of feeding livestock, value using for other household purpose in the past 12 months for each crop by-products.

The main differences in the crop expenditure compared to VHLSS2002 are the more detail in cost items and the separated cost of rice.

The section of livestock production (5B3 in VHLSS 2002 and 4B2 in VHLSS 2004) is divided in more detail about the livestock than in 2002 in both revenue and cost. However there is no question about the cost of maintenance and small repair in 2004.

In the section “forestry” (section 5B5 in VHLSS 2002 and section 4B4 in VHLSS 2004), the information for four activities (forest plantation, protection, maintenance, improvement; trees for breeding and other products collected from forests; and other forestry services; and hunting, trapping, domesticating forestry animals and birds) are collected in more detail including value of sold out, total value in 2004. The cost of this section in 2004 is collected separately for each main

activity (forestry production, forestry service and hunting activities) instead of asking for all as in 2002.

There is a minor change in the cost of aquaculture production that is the cost of medicines is combined with other expenses in 2004.

The non- farm self employment activities (section 5C in 2002 and 4C in 2004) was designed to collect additional information about these activities of the household in 2004. The code of household member who know most about this activity and the value of product currently in stock are dropped. The number of operating day per operating month, the ownership status of the activity, the percentage income from the activity that household received, the value of product from this activity is exchanged, and information about the value of by-product from this activity are added in 2004. The cost of each item was asked for each operating month instead of asking for the whole year as in 2002. In addition, the cost items are divided into more detail (14 cost items) than in 2002 (10 cost items).

The main difference in the other income section (section 5D in 2002 and 4D in 2004) is that the income from charity and support from donor or organization and others that is divided into two items that are income received from charity and support from donor or organization and income received from others.

#### Section 5. EXPENDITURE

This section is section 6 in 2002. The main difference in section 6A is the food item (chewing gum) is dropped in 2004.

In the section 5B2 (section 6B2 in 2002), the expenditure on vacation is divided into expenditure for domestic and oversea vacation.

#### Section 6. FIXED ASSETS AND DURABLE GOOD

This section is section 7 in VHLSS 2002 and it is divided into two subsections (section 6A for fixed assets and section 6B for durable good). In addition, the telephone is divided into mobile phone and fixed phone and there are two more items that collected in 2004 (micro waves and fruit grinding, machines). Three more questions are added in section 6A that are quantity of each assets, the percentage that household owned, and total amount that household spent during the past 12 months for repair and maintenance for all assets. In section 6B, the question about quantity of the asset is added.

#### Section 7. HOUSING

This is the section 8 in VHLSS 2002. There are many changes such as dropping out and adding more questions in this section compared to 2002 questionnaire. It drops the questions about previous place that household live, the material made for toilet cover, and all the questions (from question 34 to question 44) that collect information about accessing public information such as television and radio channels, and newspapers. It adds more questions about the numbers of months that household renting the house in the past 12 months, the number of months that household renting out the land or house in the past 12 months. It divides the water source into drinking and water for other uses and collects information about how frequently the household boil the drinking water.

#### Section 8. PARTICIPATION IN POVERTY REDUCTION PROGRAMS

It is the section 9 in VHLSS 2002. This is nearly new content in this section compared to 2002 questionnaire. The first question defines whether the household was classified as poor in 1999 or not and then it collects information about whether household knows about National Target Program and 135 Program and through which information channel (television, radio, newspaper

ect), the main projects from these programs, the benefit that the household got from these programs. The next questions collect information about the evaluation of the household about the changes in their living, the reason for those changes, the difficulties that they met in the past 12 months in their business. The information about the borrowing of the household is collected for both formal and informal lenders (there is only information about formal lenders in 2002) with further detail about the date, using purpose, amount paid during past 12 months for both principal and dividend and the intended date for pay back for each loan.

*Section 9. AGRICULTURE, FORESTRY AND AQUICULTURE PRODUCTION (ADDITIONAL MODULE)*

- 9.1. Agriculture, forestry, and water- surface, residential land and garden, pond next to housing land
- 9.2. Land that is rented/borrowed/temporarily exchanged
- 9.3. Land that is temporary rented out/ lent out
- 9.4. Planting
  - 9.4.1 rice
  - 9.4.2 staple food, food crops and other annual crops
  - 9.4.3 annual and perennial industrial crops
  - 9.4.4 fruit crops
  - 9.4.5. Cropping structure
  - 9.4.6 access to extension services
  - 9.4.7 conversions in agricultural, forestry and aquaculture land and over the past 10 years
  - 9.4.8. Bought, bid, inherited land or use right-transferred land over the last 10 years
  - 9.4.9. Sold, tender-expired, acquired land or inheritance right-transferred land in the last 10 years

Section 9.1, 9.2 and 9.3 is the section 5B1 in VHLSS 2002. Section 9.1 collects additional information about number of plots, the distance from the household residential to each plot, the quality, irrigating method, the ownership status, the first time that household used, the year that household has Land Use Certificate, the main household member using and managing, whether the household uses the plot for production purpose during the past 12 months. Section 9.2 collects in more detail information about rented/borrowed/temporarily exchanged land of the household in the past 12 months. It collects information about relationship with the landlord, the duration, payment method, amount of money that household paid and for how many months and the date of the latest payment. Section 9.3 collects information about land that household rented out in the past 12 months and it collects the same information as in section 9.2.

Sections 9.4.1 to 9.4.4 are mentioned in the section 5B2 in VHLSS 2002. However, they are divided in further detail for rice and other crops and these sections are mainly collected information about productivity of rice compared to previous year and the reason for the changing in the productivity of rice and who is the main trader in the past 12 months. For other crops, the main information these sections collected are about the main traders and the place that household sold the products.

Section 9.4.5 collects information about the area for each type of crop that household planted in 1993, 1998 and 2003. In additional, it collects information about the change in rice production (variety) from 1998 up to the interview time, the year using these varieties, the year stop using varieties, which information source that the household know about the variety, who they brought from and the reason they using it. The next questions collect information about the new crops that household planted since 1994, the first planted year, the information source about the new crops,

whether the household still planting it or not and the reason as well as the change in planted area for these crops.

Section 9.4.6 collects information about accessing extension services and it is the new section in VHLSS 2004. It collects information about 15 extension services including fertilizer, irrigation new seeds, land issues, weather, and livestock service, food for livestock, market information and ect... In addition, it is also collected information about the fee, the impacts of these services on decision making of the household on crop and livestock production.

Section 9.4.7 collects information about conversions in agricultural, forestry and aquaculture land during the past 10 years by types of land before and after conversion, area of conversion, the time and cost of conversion, amount received from government assistance in both cash and kind for these conversions. In addition, this section collects information about forestry planting including the area of planting, the first and the last year of planting, the cost and amount of cash and kind received from government for each plot.

Section 9.4.8 collects information about land that household bought, bid, inherited or use right-transferred over the last 10 years, including month and year that household owned, month and year that household received land certificate, type of land, area, the reason that household has this land, the cost in both cash and kind, from whom that household got the land.

Section 9.4.9 collects information about land that household sold out, tender-expired, acquired land or inheritance right-transferred land in the past 10 years. The same questions as in section 9.4.8 were used to ask the household such as the time they were sold out; area, amount received, and type of land ect... However, there are some additional questions about the reason and the compensation received for land was acquired by government, the money received and the relationship of the buyer for the land sold out.

#### Section 10. NON- FARM ACTIVITIES (ADDITIONAL MODULE)

10A. INFORMATION ABOUT TIME. LOCATION AND LABOR

10b. BUSINESS HISTORY

10C. PARTICIPATION IN BUSINESS ASSOCIATIONS AND CLUBS

10D. CONTACT WITH GOVERNMENT AGENCIES

10E. OTHER CHARACTERISTICS

10F. THE ACTIVITIES THAT OPERATED DURING THE PAST 10 YEAR AND STOPPED IN THE PAST 12 MONTHS

Section 10A collects information about whether the non- farm activities is operated in every month in the past 12 months or not and the reason for not operating in the whole year, the operating place, the total number of labor and number of paid labor.

Section 10B collects information about history of the non- farm activities, including the opening year, the founder, number of other competitors, and the relationship of the household with these competitors, the main difficulties at the beginning of this business, total amount of money invested at the beginning, from which source that household has this amount of money, experience about this business before setting up.

Section 10C collects information about whether household is the member of Vietnam Chamber of Commerce and Industries (VCCI), member of other business union, enterprise union, and the services that household received from these unions.



Section 10D collects information about number of time that tax collectors visit the household and the reason for these visits.

Section 10E collects information about the way that household buy the inputs for these non- farm activities and the place that household sells the outputs of the activities as well as the evaluation about the important of each market for these products (internal province and outside market as well as international market). This section is also collected information about the information channels that household often collect in order to set up the price for its products, information about the business performance compared to two years ago as well as information about the major constraint for these household businesses.

Section 10F is the last section of household questionnaire that collects information about the non-farm activities that household operated during the past 10 years but stopped to operate during the past 12 months, including the first year it started, number of operating years, the household member who operated it and the reason for stopping these activities.

### *2.2.2 Commune questionnaire*

There is additional section (section 9) in VHLSS 2004 commune questionnaire compared to 2002. This section collects information about Credit and saving that will be discussed in more detail later. We will only discuss the changes between VHLSS commune questionnaire 2004 and 2002 in each section.

**Section 0.** Two more questions were added in this section to collect information about the number of years living in this commune, previous position that held for each local commune leader.

**Section 1.** This section drops out the questions about the year of data collected, the number of people working in agriculture sector and adds more questions about number of household and people living in the commune by registration status, number of people migrated in and out of the commune during the past 12 months.

**Section 2.** This section collects additional information about number of household affected by natural disaster in the past 3 years, number of household classified as poor household by National Target Program and the amount money received from assistance for natural disaster relief during the past 3 years.

**Section 3.** In this section, it drops out the questions asking about total number of employee, number of employee from each 5 main factories/ enterprises or traditional businesses increased or decreased compared to 5 years ago and the average labor cost but it adds additional questions about number of labor of the commune and number of female labor of the commune working for these enterprises. In addition, it also collects further information about enterprises/ traditional businesses that labor of commune could go to work there and come back home daily and number of enterprises/ traditional businesses located within the commune, the name of traditional businesses.

**Section 4.** In VHLSS 2002, this section is only collected for the commune that agriculture is the main income source in both urban and rural communes. However, it collects information for every commune in rural area in 2004. In addition, this section collects further information about the productivity of crop compared to 5 years ago, the reason for that change, the total crop output

compared to 5 years ago, where the farmers sold their agriculture products, the main trader, the distance from commune to the market, the assistance received from the trader, the number of traders that farmers in the commune could sell their products, types of crops and total area by crops in 2003, 1998 and 1993. It is also collected information about land situation in the commune, including the irrigation by type of land, percentage of land having Land Certificate (LUC), the reason for the land without LUC, the year that the biggest LUCs are provided, the number of LUC in that year, the main reason that household would like to have LUC. There is also additional information about the free time and lack of work for each month in the past 12 months for labor by gender is collected. In the questions asking about extension service, there are additional information about percentage of farmer participated in the extension meeting, percentage of women participated in these meetings, the lack of service that these meetings could not provide, the number of times that extension official visited the farmer and whether the commune has veterinary official, what are the most difficulties that the farmer faced to.

**Section 5.** This section collects additional information about the radio station, cultural communal house, and small irrigation station in the commune. It also collects additional information about the main transportation mean, the owner status of transportation mean, cost, time and the frequency that people living in the surveyed village often use to go to market, committee house, post office, bank, small town, center of province, big cities such as Hanoi, Haiphong and Ho Chi Minh City. It also collects more information about the main goods that is sold in the daily market, main market, and the projects including infrastructure that are implemented during the past 10 years that includes the type of project, starting year, the completed year, main budget sources, the value of the project and the contribution of the commune including household contribution, number of benefited households in the commune and number of surveyed households benefited from project.

**Section 6.** There are only two more questions added in this section in 2004 asking about whether commune have kindergarten or preschool before asking in detail information about these schools in the surveyed village as in 2002.

**Section 7.** There is only one additional question asking about the transportation mean that people living in the commune often use to go to health facilities.

**Section 8.** There is no change in this section.

**Section 9.** This is a new section in commune questionnaire. It collects information about credit and saving. First, it collects information about the main saving methods that people in the commune often use and the place that the people can put their money on by type of organizations (bank, credit institution, union, private), the distance from the commune to these organizations and whether the households in the commune can borrow from these agencies, and the purpose of using the loan.

### **3 Sample**

#### **3.1 Sample design**

The series of VHLSS from 2002 to 2010 will rely on a master sample for sample selection. A master sample is a random sample of the 1999 Population Census enumeration areas. From this sample of enumeration areas, multiple samples of households can be selected for different surveys or for the different years of a rotating panel survey such as the VHLSS. The master sample used in the VHLSS is a two-stage area sample where communes are selected in the first

stage and 3 enumeration areas (EA) per commune are selected in the second stage. The communes were stratified on province and urban/rural and the sample was allocated over strata proportional to the square root of the number of households. Both communes and EAs are selected with probability proportionate to size (PPS), the size being the number of households according to Population census 1999. The surveyed households in each selected EA is selected based on the most updated list of the households in these selected EAs (3 months before the field work) and weight is used to readjust the population change. This section describes in detail the master sample and other aspects of sample design established for the VHLSS.

### *3.1.1 Target population*

The target population of the VHLSS comprises the civilian, non-institutionalized population of Viet Nam. In order to cover this target population, interviews are conducted at the household level. Because people are mobile, it is important to define clearly where people are to be enumerated to avoid double counting and under coverage. Only persons considered as permanent residents of the household are eligible for inclusion in the surveys. Persons who are considered as permanent residents of a household but are away temporarily will be included. This includes persons on vacation, temporarily in a hospital, and students living away from home during the school year. However, any housing units containing only students living away from home during the school year should not be included in the survey because data for such students would be obtained from their permanent place of residence.

*Ineligible households.* GSO has to decide on eligibility criteria for the VHLSS. Some households may be considered as ineligible for selection in the survey. One example is student housing blocks. The households consist of rooms in a dormitory, and not actual family households. They are unstable, changing substantially from year to year. The same may be true for households in other institutions (military compounds, hospitals, prisons etc).

*Ineligible EAs.* Some EAs consists entirely or to a large part of households that are not eligible for selection in the survey. EAs which consist primarily or entirely of ineligible households should, if possible, be excluded from the sample frame (for example EAs that consists almost entirely of student housing blocks).

### *3.1.2 Three-stage design*

The sample design consists of three stages with communes/wards selected at the first stage, census enumeration areas (EAs) as the secondary sampling units (SSUs) and households selected at the third stage.

#### Primary sampling units

The sampling units of the first stage of sample selection are called primary sampling units (PSUs). PSUs must be sufficiently large to allow for repeated sampling of households within them. In Vietnam there are two choices of PSUs for a master sample: communes (or their urban equivalent the district town or ward) or EAs. Communes contain on average about 1600 households while EAs were set up during the census to contain about 100 households (Table 3.1). PSUs should contain a minimum measure of size of 70 to 75 households to permit repeated sampling of households for the VHLSS or other surveys. However, about 8000 EAs (5% of all EAs) contain less than 70 households and would have to be combined which would be a time-consuming task.

Based on the above discussion, communes were chosen as PSUs, with three EAs to be selected per commune at the second stage, using only one of EA for each year of VHLSS survey. This is

technically a three-stage design (counting the selection of households), but it is operationally equivalent to a two-stage design.

This design solves several problems simultaneously:

1. The problem of the large size of communes does not arise because we are working only one of the selected EAs for each survey.
2. Second, this design allows for rotation of EAs, rather than households, which is operationally less complicated.<sup>2</sup>
3. The problem of small EAs is not as serious since it is the EAs, rather than households, that will be rotated over time. If the need arises to group EAs to satisfy a minimum measure of size, then such grouping can be restricted to only the selected communes.

Table 3.1: Summary statistics on communes/wards and enumeration areas, 1999

	Urban areas	Rural areas	All areas
Number of households (1000)	4,026 (24%)	12,635 (76%)	16,661
Number of EAs	38,435 (23%)	128,085 (77%)	166,520
Number of communes/wards	1,561 (15%)	8,915 (85%)	10,476
Households per EA	105	99	100
Households per commune/ward	2,579	1,417	1,590
Number of EAs per comm./ward	25	14	16

Source: *Census of Population and Housing 1999*

Urban wards and district towns are on average almost double the size of rural communes, giving a fairly low number of wards as compared to the number of rural communes. This is not a problem unless the sample of urban wards/district towns constitutes a large fraction of the total number of PSUs. If the fraction exceeds 50% one should reconsider using EAs rather than wards as PSUs in urban areas.

Communes as PSUs should also not be too small. A minimum size of 300 households is required for the commune to serve as a PSU because a sub-sample of three EAs is to be selected from the commune. Altogether 529 communes/wards in 38 provinces were found to have less than 300 households. Some provinces have a large proportion of small communes, for example Lao Cai rural where 50 out of 161 communes have less than 300 households. Only 4 urban wards have less than 300 households

#### Secondary sampling units

Because some communes/wards are quite large, a two-stage design with selection of households randomly within a commune/ward would have created difficulties in implementing fieldwork so a second stage of selection was introduced in which three EAs are selected per commune. However, in each survey only one of the three EAs will be used. This is technically a three-stage design (including the selection of households), but it is operationally equivalent to a two-stage design as only one EA is selected within each commune for a specific survey. This design allows for rotation of EAs, rather than households within each EA, which is operationally simpler. Also,

<sup>2</sup> As explained below, each time the VHLSS is completed half of the households will be replaced by new households. It means that a half of EAs is replaced and the surveyed households in the remaining EAs are resurveyed (panel). This replacement is referred to as rotation.

the problem of small EAs is not as serious since it is the EAs, rather than households, that will be rotated over time.

### Third stage sampling units

Households are the third stage sampling units. Sample size in terms of number of households and the number of households to be selected per EA will be discussed below.

### *3.1.3 Sample Size*

The proposed targeted sample size for the VHLSS in 2002 was 30,000 completed household interviews in order to get precise estimates at the provincial level. A much smaller sample size would have been adequate for the production of precise estimates at the national and regional levels and for urban and rural areas within these domains.

This section discusses the factors that should be taken into consideration in the determination of sample size and, in the process, provides a justification for the proposed sample size of 30,000 households.

Three major issues considered in determining the appropriate sample size for a survey were:

- the precision (reliability) of the survey estimates
- the quality of the data collected by the survey, and
- the cost in time and money of data collection, processing, and dissemination.

The following subsections discuss each of these issues in turn.

### The Precision of Survey Estimates

Understanding the concept of the precision of a sample survey estimate is critical to making a decision about sample size. By definition, a sample taken from a population is not a complete picture of the population. However, an appropriately drawn random sample of reasonable size can provide a clear picture of the characteristics of that population, certainly sufficient for decision-making purposes. From a sample of households one can collect data and generate a sample (or survey) estimate of a population parameter. The population parameter value of a characteristic of interest is generally unknown.

The question of interest is: how close can one expect a sample estimate to be to an unknown population parameter value? A survey estimate is variable because it depends on the random sample on which it is based. A different sample might yield a different estimate of the same parameter value. The question is: Just how variable is it? The variability of a survey estimate can be measured by a statistic known as the standard error of the estimate, which is defined below.

The sample design used in the Household Living Standards Survey is a three-stage stratified cluster design. This is a standard design used for household surveys worldwide. Households to be sampled are clustered within sampled areas, saving travel and administrative costs. However, clustering results in somewhat higher variability of survey estimates because households within the same cluster often are somewhat similar to each other in terms of measures being obtained from the survey questionnaire (e.g., income or education). To the extent this is true, the variability of estimates from a clustered sample is higher than what would be obtained from a simple random sample of the same number of households. A factor representing the increase in variability for a proposed sample design compared to a simple random sample is called a “*design*

*effect*”. We expect the design effect for the sample design proposed here to be somewhere between 1.5 and 2 for most variables of interest.

To simplify matters, consider the estimation of a percentage, for example, the percentage of households in Vietnam that are poor. Let  $p$  denote the estimated percentage. An equation for the calculation of the standard error of the estimated  $p$  obtained from a given complex sample design, denoted by  $SE(p)$  is given by:

$$SE(p) = \sqrt{1-f} \sqrt{deff * SRSVAR(p)}$$

where

$$\begin{aligned} f &= n/N, \\ n &= \text{the sample size of the survey,} \\ N &= \text{the total number of households in Viet Nam,} \\ deff &= \text{the design effect associated with the complex design of the survey (due to stratification, clustering, and unequal probability sampling), and} \\ SRSVAR(p) &= \frac{p(1-p)}{n}, \text{ the variance of the estimate under simple random sampling.} \end{aligned}$$

Note that  $f$  represents the proportion of the total population of households sampled. The quantity  $1-f$  is known as the finite population correction factor (fpc). For the rest of this discussion, a design effect of 2 is assumed.

Clearly, the value of  $SE(p)$  representing the variability of the estimate  $p$  depends on the following factors:

- the value of  $p$
- the size of the sample, and
- the proportion of the population being sampled.

This is illustrated in Table 3.2.

Table 3.2: Estimated Standard Errors and 95% Confidence Intervals for a survey estimate of 50% for various sample sizes (assuming a design effect of 2).

Column (1)	Column (2)	Column (3)	Column (4)	Column (5)	Column (6)
n	sqrt(1-f)	sqrt(SRSVAR(p))	SE(p)	95% LCB	95% UCB
100	1.00000	5.00000	7.07	36.14	63.86
200	0.99999	3.53553	5.00	40.20	59.80
300	0.99999	2.88675	4.08	42.00	58.00
400	0.99999	2.50000	3.54	43.07	56.93
500	0.99999	2.23607	3.16	43.80	56.20
600	0.99998	2.04124	2.89	44.34	55.66
1,000	0.99997	1.58114	2.24	45.62	54.38
5,000	0.99985	0.70711	1.00	48.04	51.96
10,000	0.99971	0.50000	0.71	48.61	51.39
15,000	0.99956	0.40825	0.58	48.87	51.13
20,000	0.99941	0.35355	0.50	49.02	50.98
30,000	0.99912	0.28868	0.41	49.20	50.80

40,000	0.99882	0.25000	0.35	49.31	50.69
50,000	0.99853	0.22361	0.32	49.38	50.62

Table 3.2 presents the estimated standard error (column 4) of a variable whose true value is 0.5 for several different sample sizes of households (column 1), each representing a different proportion of the total population of households (column 2, where the total number of households was taken to be 17,000,000). For example,  $p$  might be the percentage of households who are poor. The estimate of 50% is used in these calculations because it is the one for which the variance of the estimate under simple random sampling is highest. The square root of the variance under simple random sampling is given in column 3. The standard error of the estimate (column 4) is obtained by multiplying the values in column 2 (the square root of the fpc), the square root of the design effect (assumed to be 2), and column 3 (the square root of the variance of the estimate under simple random sampling).

The first thing to note is that the values in column 2 change by very little as the sample size increases. This shows that the impact of the finite population correction has a very minor effect on the variability of the survey estimate. As a result, in estimating the variability of estimates, the values of  $f$ , the proportion of the population not sampled, is usually ignored.

On the other hand, as the sample size increases, there is a noticeable reduction in the variability of  $p$ , as measured by its standard error. The value of  $SE(p)$  is around 7 for  $n=100$  (that is, a sample of size 100) but only around 0.3 for  $n=50,000$ . Yet it is also noteworthy that the degree of the reduction decreases as the sample size increases. This is because the value of  $SE(p)$  is a function of the square root of the sample size ( $n$ ) rather than the sample size itself.

In summary, Table 3.2 illustrates the fact that:

- i.) The proportion of the total population actually sampled has very little impact on the precision of the survey estimates even for sample sizes as large as 10 percent of the population.
- ii.) The actual sample size has a strong effect on the precision of survey estimates.
- iii.) The impact that sample size has on precision decreases as the sample size increases.

A national sample of 30,000 households will result in a confidence interval of between 49 percent and 51 percent for Vietnam, and between 44 percent and 56 percent for a typical province. Knowing that a population value is likely to be between 44 percent and 56 percent will generally be sufficient for most decision-making purposes regarding funding or other resource allocation decisions at the provincial level. The databases provided by the GSO suggest that reasonably precise estimates can be obtained for at least half of the provinces. Moreover, with the proposed sample size, reasonable precision will be obtained for survey estimates for about half of the provinces, where sample sizes will be based on roughly 500 or more households. Within many of the larger provinces, reasonable precision can be obtained separately for urban and rural domains as well. For these reasons and reasons of data quality, cost, and timelines discussed below, this report recommends a sample size of 30,000 completed household interviews.

#### Data Quality.

An important consideration in the determination of sample size is the quality of the data to be obtained from the survey. It is important to maintain data of the highest possible quality so that one can have confidence in the estimates generated from them. Checking the quality of the data at every stage of the implementation of the survey is very important. As a result, it is important to

keep the number of questions to a reasonable size so that adequate checking and editing can be done in a fashion that is efficient both in terms of time and money. This is particularly important because of the extensive nature of the questionnaire planned. In general, the longer the questionnaire, the fewer the number of respondents that should be targeted.

Another factor related to sample size that affects quality is the number of staff working on the survey. For instance, smaller sample sizes require fewer interviewers, and so these interviewers can be more selectively chosen. In particular, with a smaller sample size, it is more likely for all interviewers to be recruited from the ranks of the well-trained and experienced staff of the GSO and its provincial offices. Moreover, better training will result, since it can be better focused and proportionately more survey resources can be devoted to such training. Fewer training materials will be needed and interviewers will receive more individual attention during training and in the field. All of this will result in fewer problems in data collection in the field and in subsequent editing of the data collected. Consequently, the data available for analysis will be of a higher quality, permitting policy makers to have greater confidence in the decision being made on the basis of these data.

In addition to concerns about the quality of the data collected, larger sample sizes make it more difficult and expensive to reduce survey non-response. It is important to keep survey non-response as low as possible, in order to reduce the possibility that the survey estimates could be biased in some way by failing to include (or including a disproportionately small percentage of) a particular portion of the population. For example, persons who live in urban areas and have relatively high incomes might be less likely to participate in the survey. Failure to include a large segment of this portion of the population could affect national estimates of average household income, educational attainment, literacy, etc. With a sample size of 30,000 households, as opposed to a larger number of households, it will be much easier and more cost-effective to re-visit households that initially choose not to participate, in an attempt to persuade them to do so. Persuading initial non-participants to become participants can be a costly and time-consuming job. It is important to the quality of the survey data that adequate resources and time be made available so that a good job of what is known as “refusal conversion” can be done.

#### Cost and Timeliness

The size of the survey can dramatically affect the cost in terms of money. It also will affect the time in which the data can be made available for analysis. It is important that data and survey estimates be made available as quickly as possible to governmental agencies, so that policy decisions can be made on the latest data available. The larger the database, the longer it will take to do a good job of cleaning, editing and weighting the data for analytic purposes.

The difficulties associated with implementing surveys with large sample sizes are amply illustrated by the experience of the GSO with the 1997 MPHS with a sample size of 45,000 households. These difficulties prompted the GSO to reduce the sample size of the 1999 MPHS to 25,000 households. The 1999 survey was successfully implemented and data processing progressed smoothly. The sample design for the VHLSS is required to provide estimates for each province. The conflict is that large sample size needed to get such estimates would require a survey of unmanageable size. A total sample size of 30,000 households would be manageable and would give reasonably good estimates on provincial level when analyzing trends over an 8 to 10 year period.

#### *3.1.4 Cluster size*

There are two main alternatives for selecting cluster size. If cluster size is allowed to vary, the number of households per cluster can be selected so that the overall probability of selection



within a province is the same for all households. The alternative is to select a fixed number of households with sampled EAs. Selecting a fixed number of households per EA is easier to implement and is the procedure traditionally used by the GSO. However, this procedure produces more variation in the sampling weights, which reduces the precision of survey estimates.

In general, the optimum number of households to be selected in each PSU will depend on the cost structure of data collection and the degree of homogeneity or clustering of the survey variables within the PSU. PSUs will typically consist of households that are located near to each other. They will thus tend to be more alike in terms of important socio-economic characteristics (for example, income, education, occupation, etc.), than households that are in the same general area (for instance, in the same commune) but further apart. This increase contributes to the so-called *design effect*. In other words, the precision of survey estimates will be lower compared to a design with households that are not so tightly clustered. On the other hand, clustering has considerable cost advantages, and it is possible to “trade off” the cost savings associated with clustering against the corresponding loss in precision.

An expression for calculating the design effect (denoted as *deff*) associated with clustering is given by

$$deff = 1 + (b - 1)\rho,$$

where  $\rho$  (also written as 'roh') is the intra-cluster correlation, and  $b$  is the number of housing units to be selected in each PSU, that is, the cluster size. The design effect represents the factor by which the simple random sample variance must be multiplied to take account of the clustering in the sample design. Failure to take account of the design effect in the estimates of sampling errors can lead to invalid interpretation of the survey results. It should be noted that the magnitude of *deff* is directly related to the value of  $b$ , the sub-sample or “cluster size.” For a fixed value of  $\rho$ , the design effect increases linearly with  $b$ . Thus, to achieve low design effects, it is desirable to use as small a cluster size as possible. Table 3.3 illustrates how the cluster size and  $\rho$  affect the design effect. For example, with a cluster size of 50 housing units per PSU and a  $\rho$  of 0.02, the design effect is 1.98. In other words, this cluster sample design is only as precise as an unclustered sample of half the total number of households. With larger values of  $\rho$ , the loss in precision is even greater as can be seen in the right hand side of Table 3.3.

Table 3.3: Design Effects for selected combinations of Cluster Size ( $b$ ) and Intra-class Correlation ( $\rho$ )

Cluster Size	Intra-class Correlation								
	0.005	0.01	0.02	0.03	0.04	0.05	0.10	0.20	0.30
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.05	1.09	1.18	1.27	1.36	1.45	1.90	2.80	3.70
15	1.07	1.14	1.28	1.42	1.56	1.70	2.40	3.80	5.20
20	1.10	1.19	1.38	1.57	1.76	1.95	2.90	4.80	6.70
30	1.15	1.29	1.58	1.87	2.16	2.45	3.90	6.80	9.70
50	1.25	1.49	1.98	2.47	2.96	3.45	5.90	10.80	15.70
75	1.37	1.74	2.48	3.22	3.96	4.70	8.40	15.80	23.20
100	1.50	1.99	2.98	3.97	4.96	5.95	10.90	20.80	30.70

Methods for optimizing the size of cluster are well known (e.g., see Cochran, 1977, Chapter 9). As indicated above, the effective sample size generally decreases with increasing levels of clustering i.e., increasing values of  $b$ .

It is important to recognize that clustering is only one of three components of design effects. The other components are stratification and weighting (see Yansaneh and Eltinge, 2000, and references cited therein). Stratification generally results in design effects less than one, whereas clustering and weighting result in design effects greater than one. Any proposal regarding cluster size must take into consideration the estimated increase in design effects of about 10 percent due solely to differential sampling rates across strata.

For the VHLSS, an optimal cluster size would be 10 in urban areas and 20 in rural areas, but because of high costs of implementing the survey in a large number of clusters, the final cluster size was selected to be 25. EAs with minimum size of 70 households are of adequate size for selection of 25 households per cluster.

#### Augmentation of the sample in 2002

After design and selection of the master sample was completed, there remained a strong concern that the sample was too small and would only produce an estimated percentage at the provincial level with a confidence interval of plus or minus six percentage points. After much discussion the GSO decided to augment the sample for the VHLSS 2002 in all provinces, with the augmented sample only responding to a shortened version of the questionnaire focused on collecting income information.

The augmentation of the sample was done by doubling the number of PSUs to 700 urban wards and district towns and 2300 rural communes. In 2002, the original sample was divided into 4 parts implemented in one month of each quarter of the year (see table 3.4).

Table 3.4. Distribution of sample in 2002 by quarters

	Quarter I	Quarter II	Quarter III	Quarter IV	Total
Long questionnaire	7500	7500	7500	7500	30000
Short questionnaire	22500	22500			45000
Total	30000	30000	7500	7500	75000

#### *3.1.5 Stratification*

The purpose of stratification is to increase the precision of survey estimates. Prior to sample selection, the sampling frame needs to be divided into mutually exclusive and collectively exhaustive categories called strata. Sample selection is then carried out independently within the strata. The formation of the strata depends to a large extent on the domains of interest for the survey. In response to the demand for accurate data at the provincial level the master sample used explicit strata based on the provinces and urban/rural status within provinces.

#### Sample allocation to strata

There is a large variation in the number of households among provinces from 57,000 in Bac Kan province to over 1 million in HCMC. This leads to a conflict between demands for equal precision in estimates for all provinces (implying equal sample sizes within provinces) and the interest to get as good precision as possible in the national estimates (implying sample sizes

proportional to the population size in the provinces). A compromise solution was used in the VHLSS by selecting sample size in each provincial strata proportional to the square root of the number of households in the given strata.

Within each province, further stratification was done by urban/rural residence to ensure high level of precision in the urban and rural estimates. The number of urban and rural PSUs was allocated proportional to the number of households in urban and rural areas.

### *3.1.6 Eligibility and response rates*

Response rates in Vietnam are generally high compared to experience in developed countries. Nevertheless they must be taken into consideration in sample design. Eligibility rates must also be considered in sample design, however for the VHLSS the target population corresponds very closely to the third stage sampling unit as institutions are by definition not listed in the sampling frames. However, it is possible that some EAs may contain a large number of student houses which are not to be interviewed because students are still considered as residents of their parents' household.

To account for attrition due to non-response at the household level, reserve households should be selected in each EA. The reserve should consist of 2 or 3 households for every 10 households selected in the EA. The extra households can be kept in reserve and used to replace households that for any reason fail to participate in the survey.

### *3.1.7 Sample rotation*

The procedure for achieving a fifty percent rotation of households from one implementation of the survey to the next (every two years for the VHLSS) depends on the units used to construct the master sample. With the master sample PSU being the commune, the fifty percent rotation is implemented by retaining the EAs in half of the communes and collecting information from the same households as last survey, and selecting new EAs in the other half. Since any possible rotation of EAs takes place every two years or more, the operational and cost problems associated with rotation EAs that are far apart are minimal because there is a long lead time for the fielding of a new EA. There are other models for rotation, but it is important to avoid data collection from two EAs within a commune during the same round of the survey.

## *3.2 Sampling procedures*

Sample selection of the communes/wards and EAs was done as part of the development of the master sample. Each year that the survey is implemented, one of the EAs in each commune must be selected, the household listing updated in that EA and the sample of survey households selected. Sample selection basically consists of preparing and updating a sample frame and randomly selecting units from that frame.

### *3.2.1 Selection of the PSU*

#### *Preparation of the Commune/ward sample frame.*

The sampling frame for selection of enumeration areas in the Household Living Standards Survey is the list of all enumeration areas identified and used in the 1999 population and housing census of Vietnam.

After examining the Excel sheets listing commune name and sample size kept by the Labor and Population Department of the GSO, about 529 communes in 38 provinces were identified to have less than the minimum measure of size (300 households) required for a PSU. All the communes

with less than 300 households were combined with adjacent communes using administrative maps to ensure contiguity.

Within each province, the name and measure of size (number of households from the 1999 Population Census) in urban wards and district towns were listed separately from rural communes to allow for independent sample selection within urban and rural strata.

#### Sampling of PSUs (communes/wards/district towns)

The sampling of primary sampling units was done with systematic PPS sampling separately within each stratum (province\*urban/rural) by the Social and Environmental Statistics Department of the GSO. The communes (or combined communes) within each province are listed in order from north to south and east to west ensuring a geographical spread of the sample. The same was done in the urban stratum although in this case the listing was done within each urban location.

The next step was to compile the cumulative number of households for the PSUs on the list. The total number of households in the stratum is then divided by the number of PSUs to be selected from the stratum. The result,  $s$ , is the “step” to be used as sampling interval in the systematic selection. Next, a random number,  $a$ , between 1 and  $s$ , was selected. The first PSU selected is the PSU containing household number  $a$ . The next PSUs to be selected are the PSUs containing household number  $a+s$ ,  $a+2s$ ,  $a+3s$  and so on.

The result of the original sampling was 700 urban wards and district towns and 2300 rural communes.

### *3.2.2 Selection of secondary sampling units*

#### Preparation of EA sample frames

During census data collection, communes or wards were partitioned into Enumeration areas. Maps of relatively high quality exist for all communes/wards, and all EAs within each commune/ward. The commune/ward maps clearly identify the EAs located in them. For the selected communes/wards a list of all the EAs was prepared. All EAs with less than 70 households were combined with adjacent EAs. (with some adjustments in remote, mountainous areas).

Lists of EAs were obtained from CD-ROM discs at the Department of Population and Labor.

#### Sampling of EAs within selected communes/wards

Within each selected commune/ward three EAs were selected. The selection was done by a PPS procedure in the same way as was done when selecting communes/wards within the stratum. Numbers 1, 2 and 3 were assigned to the selected EAs according to the selection sequence. All EAs having number 1 (one EA in each selected commune/ward) constitute the sample for the first survey year (2002).

### *3.2.3 Selection of households*

#### Updating of household sample frames

The sampling frame for selection of households is a list of all households in the enumeration areas updated by provincial statistical officers visiting the communes/wards during the year prior to implementing the survey.

### Sample selection of households

In the EAs a sample of households was selected systematically with 20 households in rural EAs and 10 households in urban EAs. The selection procedure was as follows:

The total number of households in the EA were divided by the total number of households to be selected from the EA. The result,  $s$ , is the step to be used as sampling interval in the systematic selection. Next, a random number,  $a$ , between 1 and  $s$ , was selected. The random number was selected by the field supervisor from a list of random numbers provided by GSO.

The first household to be selected is the household number  $a$ . The next households to be selected are the household number  $a+s$ ,  $a+2s$ ,  $a+3s$  and so on.

#### *3.2.4 Sample selection for the augmented sample for 2002*

**Urban.** In the *urban* domain 700 EAs were selected with a sample size of 25 per EA. we have selected 175 EAs and 25 households/EA for the ordinary survey in January. For the other three survey months in 2002 we have selected 525 EAs and 10 households/EA. *The additional sample should consist of an additional 10 households in each of the 175+525 urban EAs.* This gives an additional sample of 7,000 urban households.

**Rural.** In the rural domain we have selected 287 EAs and 20 households/EA for the ordinary survey in January. For the other three survey months in 2002 we have selected 863 EAs and 20 households/EA (altogether 1150 EAs for the four survey months). It is not possible to use the ordinary EAs for the additional sample as is done for the urban domain. This would result in a sample of 40 households from each EA, which is far above the optimum. Instead, we have to select a *parallel sample* of 1,150 PSUs and to select one EA in each selected PSU. We can do this by repeating the steps for selecting the ordinary rural sample. There is in this case no need for ensuring an EA size of at least 70 households (as we have to do in the ordinary sample), it suffices to ensure a size of at least 25 households. The total of selected EAs in rural area in 2002 is 2300 EAs.

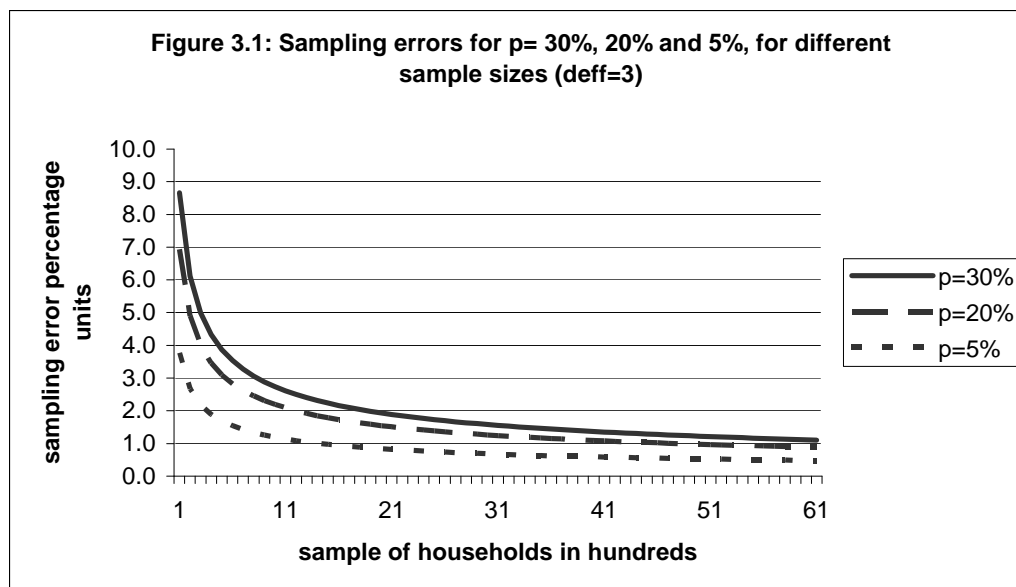
The total of sample for 2002 is 75000 households (25 households are selected to interview in each EAs). Of which there are 30,000 households are collected information of the long questionnaire.

### 3.3 Revisions to sample design for the 2004 VHLSS

The implementation of the 2002 VHLSS provided valuable information which was useful in reducing sample size for the 2004 VHLSS.

#### *3.3.1 Expected sampling errors under different assumptions on sample sizes*

Figure 3.1 shows the sampling errors for different sample sizes for the cases when the estimated proportion is 30%, 20% and 5% assuming a design effect of 3. It is apparent from the diagram that there are rapidly diminishing returns from expanding the sample above 800-1000 households.



A sample of 800-1000 households would be adequate for a single national estimate but in a national survey we always want to compare different household categories like urban/rural, education levels of head of household, household business/no household business etc. If we want to break down the results in two categories (e.g. urban/rural) we need essentially to double the sample and for five categories (e.g. household income quintiles) we need a five-fold increase in the sample size.

A sample of 10,000 households would be sufficient for a survey with the purpose of presenting national estimates for major household categories. For analysis on an individual level this would give a very large sample of approximately 50,000 persons which should be more than enough for most purposes.

There is a desire for the sample to provide estimates at a province level. The problem is to find a balance between the demands from the provinces (with a tendency towards large samples) and the interest of GSO to keep the sample on a manageable level and within the budget allocated. An important function of the survey at the provincial level is to provide estimates of change for the monitoring of poverty rates.

Figure 3.2 shows the probability to detect a change in a proportion (e.g. poverty rate) between two survey years. Two cases are presented:

- the true (unknown) change is high: we assume that the change = 8 percentage units
- the true change is “average”: we assume that the change = 4 percentage units

We have assumed that the proportion is 32% in the base year and 24% or 28% in the follow-up year. Further, we assume that 50% of the households from the base year sample are retained in the follow-up year sample. The design effect is again assumed to be 3.

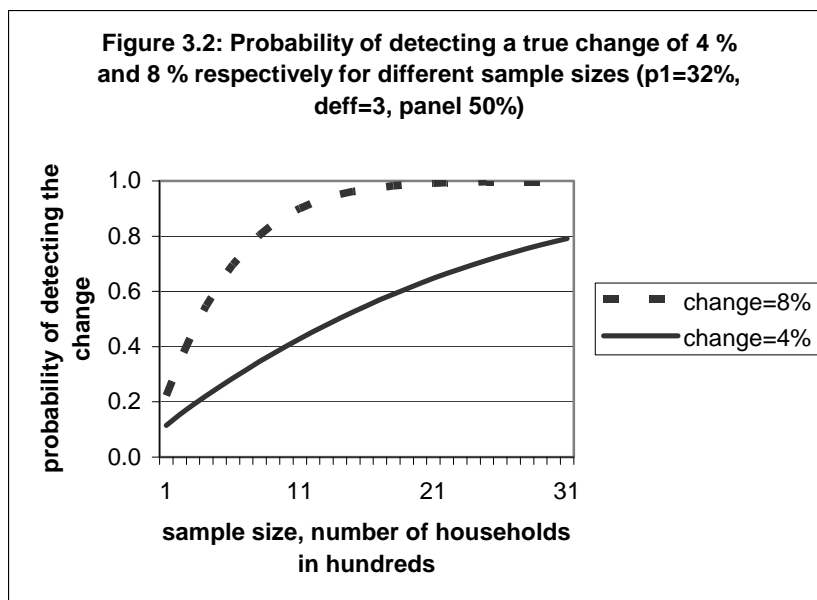


Figure 3.2 shows that when the true change is 8% we would have a 75% probability of detecting this change with a sample of 500 households. When the true change is 4% (a “normal” change over a two-year period) we would need a sample of approximately 2700 households to detect the change with 75% probability. It is apparent that even with very large samples in the provinces we would not be able to detect anything other than large changes in the rate over a two-year period. Also, the true, unknown, rate could be subject to “random” fluctuations over a short time if a large number of households are close to the poverty line. This makes it difficult to draw valid inferences from short-term changes.

The situation is much better if we look at a six- or eight-year period. Over such a long period any short-term random fluctuation of the true rate would be small compared to a steady decline of the true rate over time. Furthermore, a 2% annual change over a six-year period gives an overall change above 10%, a change that would be detected even with a 500 household sample. When we look at a longer period we also have the opportunity to pool neighboring rounds in order to gain sample size, rounds 2002+2004 could for example be compared to rounds 2008+2010. In this case we would have doubled the sample sizes and we would estimate changes over a six-year period where we would expect substantial change in poverty rates.

### 3.3.2 Expected sampling errors at provincial level for different sample sizes

The sampling errors depend on the total number of households in the sample but also (through the design effect) on the number of sampled households per PSU (enumeration area). Reducing the total sample by reducing the number of PSUs increases the sampling error much more than reducing the number of households/PSU. Table 3.5 presents some alternative sample sizes at provincial level. The increase in sampling error is rather small when going from the VHLSS-02 sample size in alternative 3 (=1200 households per province on average) to the much smaller sample in alternative 7 (=500 households). This is partly because the reduction has been done entirely within PSUs by reducing the number of households/PSU from 25 to 10 but keeping the number of PSUs unchanged (=50).

In alternative 6, on the other hand, the number of PSUs have been reduced from 50 to 25. This has a much stronger effect on the sampling error. The sampling error is substantially larger in alternative 6 as compared to alternative 7 but the total sample size is the same in both alternatives. If we want to see what happens if we *expand* the VHLSS-02 sample from 75,000 households to 90,000 households we should compare alternative 2 and 3. In alternative 2 we have assumed that the sample size within enumeration areas (PSU) is kept at 25 households but the number of PSUs has been increased to 60 per province, approximately 3700 PSUs in total. It is apparent that such an expansion doesn't improve things very much. The slight improvement in precision will come at a high cost; the field work costs for alternative 2 should be 20% higher than the costs for alternative 3.

Table 3.5: Expected sampling errors and confidence intervals for different assumptions on number of PSUs (enumeration areas) and number of households per PSU in the sample

No. of Hh/prov. (total sample size)	No of PSUs	hh/ PSU	deff w	deff cl	p=5 s.e. (95% CI)	p=10 s.e. (95% CI)	p=15 s.e. (95% CI)	p=30 s.e. (95% CI)
1. 2500 (150000)	100	25	1.1	3.3	0.8 [3.3-6.7]	1.2 [7.7-12.3]	1.4 [12.3-17.7]	1.8 [26.5-33.5]
2. 1500 (91000)	60	25	1,1	3,4	1,1 [2,9-7.1]	1,5 [7,1-12.9]	1,8 [11,5-18.5]	2,3 [25,5-34.5]
3. 1250 (75000)	50	25	1.1	3.3	1.2 [2.7-7.3]	1.7 [6.8-13.2]	2.0 [11.2-18.8]	2.5 [25.1-34.9]
4. 1000 (60000)	50	20	1.1	2.9	1.2 [2.6-7.4]	1.7 [6.7-13.3]	2.0 [11.0-19.0]	2.6 [24.9-35.1]
5. 750 (45000)	50	15	1.1	2.4	1.3 [2.5-7.5]	1.8 [6.5-13.5]	2.1 [10.8-19.2]	2.7 [24.7-35.3]
6. 500 (30000)	25	20	1.1	2.9	1.7 [1.6-8.4]	2.4 [5.3-14.7]	2.9 [9.4-20.6]	3.7 [22.8-37.2]
7. 500 (30000)	50	10	1.1	1.9	1.4 [2.2-7.8]	1.9 [6.2-13.8]	2.3 [10.5-19.5]	3.0 [24.2-35.8]
8. 375 (23000)	25	15	1.1	2.4	1.8 [1.4-8.6]	2.5 [5.1-14.9]	3.0 [9.1-20.9]	3.8 [22.5-37.5]



*Note: The design effect due to unequal weights is assumed to be 1.1. The roh is assumed to be 0.1. The design effect due to clustering (deff cl) is calculated from roh and the number of hh per PSU ( $D_{cl}^2(\bar{y}) = 1 + roh(n-1)$ ).*

### 3.3.3 Changes to sample size for the 2004 survey

Results of analysis of the 2002 survey results indicate that a sample with the ability to detect rather small short term changes at the provincial level must be very large, so large that it would be very difficult to manage. A sample of, say, 3-4000 households in the province would require hiring of temporary staff which is likely to create problems with the quality of the field work and the quality of the data reducing the value of the survey.

In the 2002 survey the sample take within enumeration areas (cluster size) was 25 households. The calculations above show that there is very little information gain from the last households in the sample from the enumeration area. The gain in precision we get from going from 15 to 25 households is small.

Based on these findings, the 2004 VHLSS survey sample design was revised as follows:

- *Sample size within PSUs (cluster sizes).* Cluster sizes were reduced to 15 households per enumeration area without any expected serious effects on the precision.
- *Sample size for the expenditure module.* The sample size for the expenditure module was reduced to 9,000 households allowing for comparisons over major groups of households and individuals, but not comparison at the provincial level.
- *Sample size for the income module.* The sample size was reduced to 36,000 households by reducing the number of households sampled per PSU from 25 to 15, not by reducing the number of PSUs.
- *Ability to detect changes in poverty rates.* The proposed reductions in sample size had only marginally reduced the probability of detecting changes in the poverty rates as the ability to detect small changes at the provincial level were already very small with the larger 2002 sample size.
- *Number of EAs in sample:* For the 2004 survey altogether 3063 EAs (63 new EAs are additional selected for 6 new provinces) were included in the sample. 1546 EAs were selected as a sub-sample from the VHLSS02 sample. The remaining 1517 EAs were selected with systematic PPS from 1517 communes. There are additional EA was selected in 2004 for 6 new provinces so that the total EA is 3063.
- *Sub-sampling from the VHLSS02 sample:* For the “old” EAs (sub-sample from the VHLSS02 sample) the households were selected as a sub-sample of the VHLSS02 households.
- *Sampling from new EAs:* For the “new” EAs the following procedure was used. Within selected EAs all households were listed and 20 households were selected with systematic equal probability sampling. A sub-sample of 15 households was selected systematically from the 20 households. This sub-sample was used for the survey and the remaining 5 households were designated as reserve households (to be used as substitutes for non-response households). From the 15 households a sample of 3 households were selected randomly by systematic sampling. This sub-sample was used for the expenditure module of the survey.

### 3.3.4 Differences between the 2002 and 2004 sample design

The sample design for the 2004 survey is in many ways better than the design for the 2002 survey. The cluster sizes for the expenditure module are smaller, 3 households per EA as compared to 5 or 20 households per EA in the 2002 survey. The design effect due to clustering ( $d_{cl}^2(\bar{y})$ ) will consequently be smaller in the 2004 survey.

The selection of a second stage sample in EAs of either 5 or 20 households in the 2002 survey resulted in a large variation in the sampling weights. The variation in the weights increases the variances as compared to a self-weighting design. The increase in the variances due to variation in the weights can be measured by the design effect due to weighting ( $d_w^2(\bar{y})$ ).

Table 3.6 shows the average design effects for the two surveys. The design effect could be looked upon as the “cost” in terms of increased variance we have from the design (as compared to a simple random sampling design). It is apparent that the “cost” for variation in the weights in the 2002 survey is high. As mentioned above, this is due to the varying sub-sample sizes (either 5 or 20 households) in the 2002 survey.

Table 3.6: Average design effects due to weighting and clustering in the two surveys

	Design effect due to:		
	Weighting	Clustering	Total
VHLSS2002	1.85	1.78	3.29
VHLSS2004	1.12	1.20	1.34

The *effective* sample size is defined as  $n/\text{deff}$ . The effective sample sizes for the two surveys (expenditure module) are:

2002 survey:  $29\,530/3.29 = 8\,975$

2004 survey:  $9\,189/1.34 = 6\,857$

The effective sample size is the “real value” of the sample in terms of a simple random sample. It is interesting to note that if the 2002 survey had used sub-sampling of an equal number of households in the EA (as the 2004 survey) the design effect would have been  $1.12 * 1.78 = 1.99$  and the effective sample size would have been  $= 14\,839$ .

From the effective sample sizes we can calculate the expected ratio between coefficients of variation (CV) for the two surveys:

$$\frac{CV_{2004}}{CV_{2002}} = \sqrt{\frac{8975}{6857}} = 1.14$$

We could expect the sampling errors to be only 14 % above the sampling errors for the 2002 survey. This is a very small difference considering that the sample size in 2004 was less than one third of the 2002 sample. (However, this favorable result is partly due to the less efficient design of the 2002 sample.)

### 3.4 Sampling evaluation

#### 3.4.1. Change in definition of households

The “demarcation” of households is another problem. There is a slight difference between the 2002 and 2004 questionnaires. In the 2004 survey the questions on household members are followed by a question on how many months the person has been staying in the dwelling during the last 12 months and a question on register for residency. In the 2002 survey these two questions were not asked. Errors in household sizes affect poverty calculations. Crude calculations show that if the average household size contains a 5 % under-reporting error the poverty rate will be biased downwards by approximately 3 percentage units.

#### 3.4.2 Comparability of sample population estimates to population projections

Table 3.7 shows estimated number of households and population from the surveys 2002 and 2004 and population projections based on census data from 1999.

Table 3.7: Comparison of survey estimates and population projections

Year	Number of households (Million)	Population (Million)	Population according to population projections (constant fertility variant)
	Standard error in brackets	Standard error in brackets	
1999	16.7	76.3	76.3
2002	17.28 (0.02)	76.6 (0.34)	79.7
2004	17.34 (0.02 )	75.5 (0.23)	81.5

The total number of households and the population in 2004 seems low. The annual growth of the population during the period is 1.3 % according to the population projections (1.2 % with the alternative “declining fertility rates”). Assuming a slowly declining average household size we should expect the growth rate of households to exceed that rate somewhat. The survey, however, shows an annual growth rate of 0.8 % between 1999 and 2004. There is possibly an under-listing of households in the selected EAs.

The estimates of total population are low compared to the census projections for both surveys but especially for the 2004 survey. This is partly due to the fact that the target population was confined to civilian persons residing in households, thus excluding persons living in institutions and military personnel living in barracks. There may also be effects of either too low household estimates or under-enumeration of household members or both.

#### 3.4.3 Design effects and rates of homogeneity

Sampling errors and design effects from the 2002 survey were utilized for planning of the 2004 sample design.

The design effect contains effects of unequal weighting and clustering. The design effects due to weighting is calculated as:

$$D_w^2(\bar{y}) = \frac{n \cdot \sum w_i^2}{(\sum w_i)^2}$$

$w_i$  is the sampling weight (household weight) for household  $i$ .

The design effects due to clustering is calculated as :

$$D_{cl}^2(\bar{y}) = \frac{D^2(\bar{y})}{D_w^2(\bar{y})}$$

The design effects due to clustering are directly related to the size of the clusters. A more portable measure is the rate of homogeneity (*roh*) which is a “synthetic” measure of the correlation within clusters.

$$roh = \frac{D_{cl}^2(\bar{y}) - 1}{\bar{n} - 1}$$

where  $\bar{n}$  is the cluster size, the number of households selected in the enumeration area.

The cluster size is 25 households for the income module. For the expenditure module there are two cluster sizes; 25 households (in 750 PSUs<sup>3</sup>), and 5 households (in 2250 PSUs). The widely differing cluster sizes cause problems in the calculation of design effects. The sample has therefore been split in two parts:

- PSUs with a sample of 5 households. The sample consists of PSUs having only 5 households but also the PSUs having 25 households, in which case 5 households were selected out of the 25.
- PSUs with a sample of 20 households. These are the PSUs originally having 25 households. However, 5 households were selected to form a separate cluster and were included in the first part above.

The reason for this 'somewhat peculiar procedure was that there was not sufficient information in the data set at hand to easily separate the 5-household clusters from the 25-household clusters.

The results of the calculations are shown in Table 3.8.

Table 3.8: Design effects and rates of homogeneity for proportion of poor households and per capita income.

Estimate	$D^2(\bar{y})$	$D_w^2(\bar{y})$	$D_{cl}^2(\bar{y})$	roh
Proportion poor households, cluster size: 5 households	2.2	1.4	1.6	0.15
Proportion poor households, cluster size: 20 households	5.0	1.3	3.9	0.15
Income per capita, cluster size: 25 households	3.5	1.1	3.1	0.09
Income per capita, cluster size: 25 households, urban	4.0	1.1	3.5	0.11
Income per capita, cluster size: 25 households, rural	3.5	1.1	3.1	0.09

<sup>3</sup> In this report we consider enumeration areas to be primary sampling units (PSU) although, in a strict sense, communes are PSUs and enumeration areas (ea) are secondary sampling units (ssu). However, when only one ea is selected from each commune we could look upon the design as a twostage design where eas serve as PSUs.

As stated above, the sampling errors are somewhat overestimated because the probable variance reducing effects due to systematic sampling have not been taken into consideration. Consequently, the *roh* values are also slightly overestimated. A *roh* value of 0.10 was used for the calculation of expected sampling errors under different sample sizes to inform the design of the 2004 Household Living Standard Survey sample design.

### 3.5 Weights and standard errors for the VHLSS sample

#### 3.5.1 Calculation of sampling weights

In order to facilitate the calculation of sampling weights, an Excel file was created which recorded the relevant information and calculated the probabilities of selection at each stage. From this spreadsheet the sampling weights were calculated.

##### First stage sampling weight

$$w_{1hi} = \frac{A_h}{n_h A_{hi}}$$

$A_h$  = number of households in stratum  $h$  according to census

$A_{hi}$  = number of households in PSU  $i$  in stratum  $h$  according to census

$n_h$  = number of PSUs selected in stratum  $h$

Second stage sampling weight:

$$w_{2hij} = \frac{A_{hi}}{A_{hij}}$$

$A_{hi}$  = number of households in PSU  $i$  in stratum  $h$  according to census

$A_{hij}$  = number of households in EA  $j$  in PSU  $i$  in stratum  $h$  according to census

Third stage sampling weight:

$$w_{3hij} = \frac{M_{hij}}{m_{hij}}$$

$M_{hij}$  = number of households in EA  $j$  in PSU  $i$  in stratum  $h$  according to survey listing

$m_{hij}$  = number of households in the sample from EA  $j$  in PSU  $i$  in stratum  $h$

The final weights, to be applied to each household in the data set, become:

$$w_{hij} = \frac{A_h}{n_h \cdot A_{hi}} \cdot \frac{A_{hi}}{A_{hij}} \cdot \frac{M_{hij}}{m_{hij}}$$

##### Assessment of sampling weights

First stage weights for approximately 100 communes are less than 1.00. This happens when the measure of size for the commune (number of households) exceeds the sampling interval (=the number of households in the stratum divided by the number of communes selected in the stratum).

The best way to handle these large communes (larger than  $M_h/n_h$ ) would have been to set them aside in a special stratum before selection and to select all communes in this stratum with certainty (probability=1).

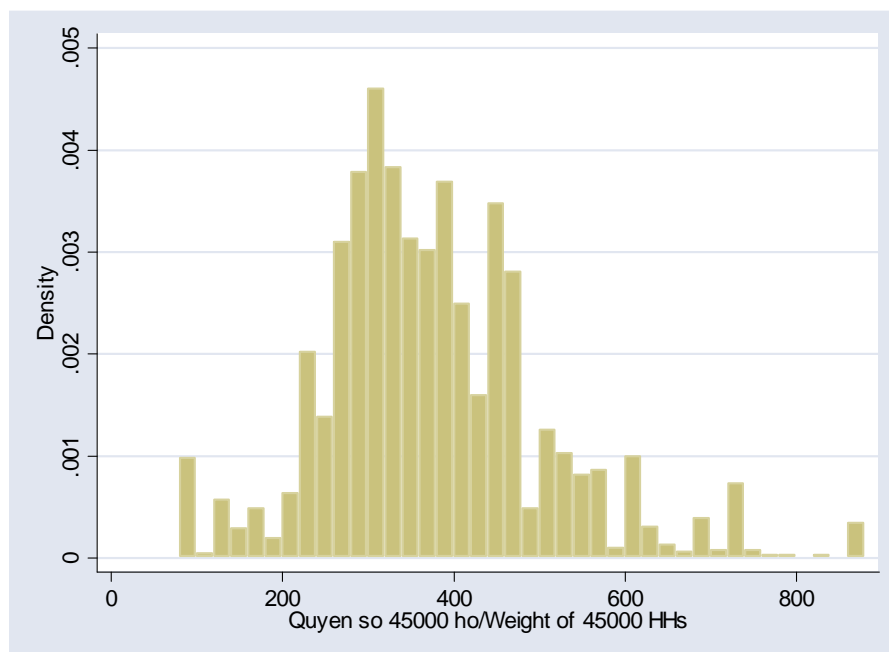
Some of the large communes have a first stage weight that is less than 0.5. In that case the commune will be selected twice when the sampling interval is applied. This will give a sample of two EAs in the commune in the second stage selection. It appears, however, that in these large communes only one EA has been selected. Whether this was deliberately done or a result of error in the sample selection has not been investigated.

Measures of size from the 1999 Census are missing for 62 communes in the file weights04. These are cases where the commune has been formed by a split from one or several communes since 1999. The missing values mean that a proper PPS selection of these communes could not be done. The best procedure would have been to distribute the 1999 census number in the old commune(s) between the new communes as good as possible in order to get a measure of size of reasonable quality also for the new communes. This seems not to have been done.

#### Sampling weights for 2004 survey

The sampling weights for the large sample (45 000 households) are shown in Figure 3.4. The group of very high weights at the right end of the histogram is the weights for urban EAs in province 701 (HCMC). The weights, from 881 to 887 reflect the fact that the sampling fraction is very low in HCMC (due to the square root allocation of the sample to provinces).

**Figure 3.4: Sampling weights VHLSS 2004**



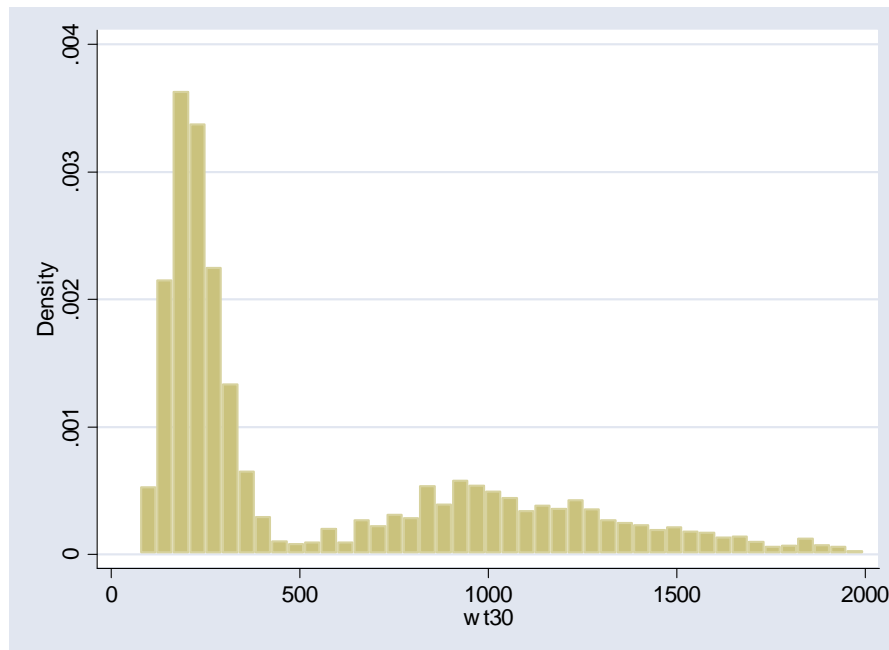
The lowest group of weights is for rural EAs in provinces 301 and 606. The allocation of sample communes to these provinces is higher than what the square root allocation will give. Example: in province 301-rural, the square root allocation calls for 18 communes to be selected from this stratum but the actual sample was 30. The province 301 has been split into two provinces between 2002 and 2003; the two parts seem both to have kept the original sample size, resulting in heavy over-sampling especially in 301 rural.

#### Sampling weights for 2002 survey.

Figure 3.5 shows the sampling weights for the 30 000 households interviewed for the expenditure module in the 2002 survey. Weights above 2000 have been excluded in the diagram (315 households had weights above 2000).

The bimodal appearance of the histogram is due to the two alternatives of sub-sampling of households in the EAs (5 households or 20 households). It is obvious from the histogram that this way of selecting the sample results in large variation in the sampling weights. This large variation in the sampling weights will unduly inflate the variances of the estimates, it would have been better to use a design with less variation in the sample take within EA.

**Figure 3.5: Sampling weights VHLSS 2002**



#### Using weight for panel data

There are about a half of households in VHLSS 2002 that are resurveyed in VHLSS 2004. The question is how to use the weight for these households when we would like to estimate for panel data? As suggestion of Mr. Hans Petterson that designed the sampling for VHLSS, it is better to use the weight of VHLSS 2004 for panel data as we have a most updated about population as well as a new sampling procedure for these households (each EA has 15 households in 2004 instead of 25 or 20 in 2002, of which 3 households are surveyed for income and expenditure questionnaire).

### *3.5.2 Calculation of standard error*

#### Standard error at national level in 2002

Sampling errors and design effects for national level estimates of poverty and per capita income are shown in table 3.9 and 3.10. The sampling errors have been calculated using the Taylor linearization estimation method for Standard errors in Stata software. The possible sampling error reducing effects from systematic sampling within provinces have not been considered when the sampling errors have been calculated. The sampling errors are therefore slightly over-estimated in the tables below.

The sampling errors are generally low, as can be expected from a sample with such large sample size. The fairly high coefficient of variation ( $cv = 8.9\%$ ) for the proportion poor households in urban is also reasonable given the low proportion.

Tables 3.9 and 3.10 also contain design effects. The design effect is the ratio of the variance calculated from the actual sample design ( $Var_{design}(\bar{y})$ ) and a hypothetical variance calculated under the assumption that the sample was selected using simple random sampling of households ( $Var_{srs}(\bar{y})$ ).

$$D^2(\bar{y}) = \frac{Var_{design}(\bar{y})}{Var_{srs}(\bar{y})}$$

The design effects are slightly lower for the estimates of proportion of poor households than for the estimates of per capita income. The smaller average cluster size in the expenditure survey (5 or 25, with an average of 10 households) as compared to the income survey (25 households) will tend to reduce the design effects.

Table 3.9: Proportion of poor households, 2002 (%)

	Estimate	Std. Err.	cv %	Deff	Obs
Total	25.5	0.4	1.7	3.0	29,513
Urban	5.2	0.5	8.9	3.0	6,909
Rural	32.0	0.6	1.7	3.2	22,604

Table 3.10: Per capita income, 2002 (household average)

	Estimate	Std. Err.	cv%	Deff	obs
Total	379.8	3.6	0.9	3.5	74,344
urban	605.9	11.6	1.9	4.0	16,026
rural	319.3	3.3	1.0	3.5	58,318

#### Sampling errors at the provincial level in 2002

Tables 3.11 and 3.12 show the median values for sampling errors, coefficients of variation and design effects for the provinces. The median sampling errors and coefficients of variation are of course much higher than on the national level due to the smaller sample sizes. The design effects remain largely unchanged as expected.

Table 3.11: Proportion of poor households, 2002 (median values for the provinces)

	Estimate	Std. Err.	cv %	Deff
Total	28.6	3.6	12.5	3.0
urban	5.7	3.4	59.0	2.0
rural	32.3	4.2	13.1	3.3



Table 3.12: Average per capita income, 2002 (median values for the provinces)

	Estimate	Std. Err.	cv %	Deff
Total	309.3	16.3	4.8	3.4
urban	463.8	49.2	10.0	3.6
rural	268.9	14.3	5.4	4.1

Table 3.13 shows the 10 provinces with the smallest sampling error and the 10 provinces with the largest sampling error in the estimates of proportion of poor households. The provinces with the smallest sampling errors all have small proportions of poor households and many of them have rather large sample sizes. For the provinces with the largest sampling errors it is the opposite situation. The table is not so informative because it mainly reflects the fact that a small proportion will have a smaller sampling error than a larger proportion when the sample size is the same (when we are discussing proportions generally below 50 % as is the case here). If the sole purpose of the survey would be to estimate changes in poverty rates we would want to expand the sample sizes in the poorer provinces.

Table 3.13: Proportion poor households (%) by province

	Tinh	Estimate	Std. Err.	Deff	Obs
10 provinces with smallest standard error					
HCMC	701	1.1	0.5	2.0	775
Da Nang	501	3.5	1.0	1.0	320
Hanoi	101	4.2	1.6	5.0	740
Ba Ria-Vung Tau	717	6.0	1.7	2.1	400
Nha Trang	511	7.7	1.8	2.2	460
Dong Nai	713	9.7	1.8	2.4	610
Quang Ninh	225	6.7	1.9	2.7	460
Bac Ninh	106	11.9	2.0	1.8	470
Ben Tre	811	14.3	2.2	2.0	500
Hai Phong	103	12.0	2.2	2.8	610
10 provinces with the largest standard error					
Gia Lai	603	56.0	4.5	3.7	460
Ha Nam	111	29.8	4.5	4.2	440
Quang Binh	407	33.1	4.5	3.9	430
Tuyen Quang	211	36.3	4.5	3.0	340
Yen Bai	213	36.8	5.2	4.5	390
Cao Bang	203	54.8	5.4	4.0	340
Lao Cai	205	49.1	5.4	4.0	340
Ha Giang	201	61.0	6.2	4.8	300
Bac Kan	207	64.4	6.9	5.8	280
Kon Tum	601	38.4	7.0	4.6	220

## **4 Implementation**

### **4.1. Central level**

Director of Social and Environment Statistics Department of GSO has overall responsibility to implement this survey. This department has responsible to organize, manage and monitor all stages of survey implementation including data need assessment, questionnaires and manual design, sampling design, training for trainers from provincial statistics offices, monitoring the training for interviewers, team leaders and field work. In addition, this department has also responsible for checking quality of data, data cleaning, analysis and dissemination.

### **4.2. Provincial level**

Director of provincial statistics office (PSO) has responsible to organize, implement the survey within his province, including personnel arrangement, training for interviewers and team leaders, field work, monitoring the implementation of the teams, data entry, data checking and cleaning. In addition, he has responsible for the quality of the data. At the beginning of each round, it is recommendation that provincial statistics office should send the supervisors to the field in order to supervise and comment for the team about the problems that they meet at the field.

### **4.3. District level**

Director of statistics office at district has responsible for implementing the survey for selected commune within his district. Each district has one team including from 3 to 4 people and team leader often is the director or deputy director of the statistics office at district.

### **4.4. Commune/ Ward level**

Leaders of selected communes or wards have the responsibility to cooperate with the survey team in order to complete their work with high quality. The commune or ward leaders often organize the meeting with the selected households to introduce about the purpose of the survey and the responsible of the household according to Statistics Law in order to get the good cooperation with the interviewer during the interview time. For the remote and mountainous areas, this work is done by village leaders. The commune/ ward leaders and team leader of the survey has setup a workplan for the field so that they could arrange the meeting with the respondents.

### **4.5. Recruitment interviewers and team leaders**

This survey is considered as the most complicated survey in Vietnam so the quality of the interviewers and team leaders is very important factor to contribute on the quality and the success of the survey. GSO recruits these people based on the education, experience, capacity in public relation, enthusiastic and good health. Each district has one team included from 2 to 3 interviewers and a team leader.

### **4.6. Training of the trainers**

Two 10 days training courses for the trainers (one in the North and other in the South) are implemented for one leader and two staffs from social and environment statistics unit of PSO. The trainers come from Social and Environment Statistics Department of GSO who are the designers of questionnaires, manual and sampling for this survey. During the training time, there

is one day for student to implement the pilot interview and the exam is given to the participants at the end of each training course to evaluate the quality of each potential trainer.

#### 4.7 Training interviewers and team leaders

The participants from above training courses are the trainers for their province. Each province has the same training course with the same content as the training course for trainers. The participants of this training course are interviewers, team leaders, supervisors.

#### 4.8. Media campaign

GSO sent to each selected household a letter to introduce the objectives of the survey and the importance of the information that the household provides to the interviewer as well as the responsibility of the household in terms of providing this information according to Statistical Law. In addition, in each center of commune or village (cultural house), GSO posted a poster about the survey to increase the awareness of the public.

#### 4.9. Field work

The field work is done about one month for each round of the survey. In VHLSS 2002, it is implemented in 4 quarters (starting at first month of each quarter) and each long questionnaire is completed within 1 and a half day and the short questionnaire is completed within 1 day. In 2004, the survey is implemented in May and September. However, most of the first round is implemented in June and some provinces completed the field for second round in late 2004 (Ho Chi Minh city).

	<b>2002</b>		<b>2004</b>	
	<i>Long questionnaire</i>	<i>Short questionnaire</i>	<i>Long questionnaire</i>	<i>Short questionnaire</i>
No of days to complete a questionnaire	1.5	1.0	2.5	1.5
<b>Field work</b>				
January				
February				
March				
April				
May				
June				
July				
August				
September				
October				
November				
December				

The long questionnaire (including two special modules) is often completed within 2.5 days and the short is completed in one and a half day. In order to implement this survey at the field, there are about 1500 interviewers, 600 team leaders, 200 provincial supervisors and 20 central supervisors

that are mobilized. GSO has also to recruit a certain number of reserved staffs for this survey to solve the cases that some survey members could not work for the survey due to health or other problems

## **5 Data processing**

The data is first checked by team leader before sending to PSO. The PSO checks the data again and send it to data entry staffs. The data entry staffs are trained about the data entry program (using CsPro) by Center of Information and Statistics. In this program, there is the program for data cleaning, it gives a warning for inconsistent or out of range for each suspicious number in each question. Based on these warnings, the PSO checks whether the errors are made by computer operator or by interviewer. For the errors made by interviewers, PSO sends them to district and asks interviewers to check the errors with households and then send the revised number back to PSO for computer operators.

After the data entry is completed at PSO, it is sent to Center of Information and Statistics as well as to Social and Environment Statistics Department. The second data cleaning process is done at the Social and Environment Statistics Department. The main errors at this stage are the inconsistent error and out of range errors. There are about 5 or 6 people (each in-charge in one or two sections) to look at the data and find out the errors. These errors again are sent to PSO for verifying and then the verified data is sent back to Social and Environment Statistics Department for the final checking. Due to it is a big survey, total time for data cleaning often takes from 8 to 12 months.

## **6 Dissemination policy**

### **6.1. Principles in VHLSS Data Dissemination**

The policy of VHLSS data dissemination should follow a number of principles to ensure the legality of data dissemination, and the effectiveness of data use.

#### *6.1.1. Legality of data dissemination*

Dissemination of statistical information is one of important functions of the GSO. As decreed in Article 1 of Degree 23/CP dated March 23, 1994 by the Government regarding the functions, duties, powers and organizational system of the GSO, the GSO is an agency under direct authority of the Government, with the function of State management in the field of statistics and provision of quantitative information on the socioeconomic situation to every agency, organization and individual in accordance with the Government's regulations. The data and information collected from VHLSS will be made available to public in due time as regulated by present and future legal documents.

#### *6.1.2. Equitability in data dissemination*

In collecting, analyzing, and disseminating data, GSO adheres to the duties and powers of an agency of the Government that have evolved to protect the impartiality and credibility of federal statistical efforts. The GSO strives for equitable policies and practices on data dissemination, ensuring that VHLSS results are revealed to all potential users including individuals and organizations both domestic and international. Everyone who meets the conditions specified clearly in the data dissemination policy can obtain the data they need. The GSO tries the best to

make the VHLSS available in an open environment, with full documentation of sampling method, questionnaire and data use introduction, summary results, and micro-data.

#### *6.1.3. Maintenance of confidentiality*

The Law on Statistics of Vietnam that was passed by the National Assembly and go to effect since January 1, 2004 requires that confidential data be not disseminated to public. According to the Law, the confidential information includes: (i) Information on identification of individuals and organizations such as name and address. These identification information can be disseminated only in case of the respondents' agreement. (ii) Confidential information of the State.

Dissemination of the summary results from the VHLSS does not have a problem of the maintenance of confidentiality because the socioeconomic indicators are reported at the regional and provincial level. However when revealing the micro-data, one should be cautious about the confidentiality of respondents. VHLSS contains a wide range information on living standards including sensitive personal information. The disclosure of sensitive information can have harmful impacts on respondent households, or even on communes. Although the identification information are deleted before micro-data dissemination, it is possible that the analysis of micro-data can locate groups of households, and especially communes. This is a central challenge to GSO to implement the dual mandates of maximizing the availability of micro-data while protecting confidentiality. Finding ways to make VHLSS data available in sufficient detail for analytic purposes will often mean stretching the limits of data dissemination up to – but not beyond – the point where confidentiality is jeopardized.

#### *6.1.4. Guarantee of data quality*

According to the Law on Statistics the statistical information disseminated by the GSO have the highest legal effects. The GSO is fully responsible for objectivity and accuracy of the collected information. This puts a pressure on the GSO in ensuring the quality of VHLSS because this is large-scale survey with a large number of questions on living standards. Although the VHLSS are implemented using extensive quality control procedures such as scientific questionnaire design, decentralized organization of fieldwork, enhanced interview skills, sound sampling method and data management, the evaluation of data quality prior dissemination is found very necessary. Analytical results of the VHLSS should be compared with results obtained from other survey conducted by the GSO and information sources of different ministries. Preliminary results of the VHLSS should also be presented to various ministries and international cooperation agencies for comments on the reliability of the data before the completed results are released to the public.

### **6.2. The Policy of VHLSS Dissemination**

Data and information collected from VHLSS are very useful for the monitoring and evaluation of living standards of people, assessment and design of socioeconomic policies and programs. As a State agency in the statistical system, the GSO has the obligation to maximize availability of

VHLSS to the public, minimizing the time from data collection to dissemination so as to maximize the usefulness of the VHLSS. At the same time the data dissemination has to ensure the confidentiality and follow the Government's regulations. Final decisions about VHLSS dissemination will depend on the specific data collected in each survey, and can only be made in detail after the data have been collected, processed, and reviewed for the confidentiality and legal framework of data dissemination issues. The policy of data dissemination proposed just provides general guidelines for the dissemination of VHLSS.

#### *6.2.1. Data to Be Disseminated*

The GSO aims to make the data and information from VHLSS available in user-friendly formats for all sorts of users, subject only to limits imposed by data quality, legal regulations, and the need to protect confidentiality. The following documents and data will be disseminated for users depending on their need.

- (1) Basic documents of VHLSS: They include the questionnaires, format files, data dictionaries, code books, brief descriptions of any constructed variables made available and brief sampling information. For some users, more detail will be required, such as the manuals for the interviewers, supervisors or data entry operators, and other information related to the survey design and data collection.

These documents are prepared in printout and electronic formats (CD-ROM or diskette) depending on the users' requirement.

- (2) Preliminary reports of VHLSS results: Information collected from VHLSS are analyzed to produce statistical abstracts and tabulation tables on general socioeconomic indicators: poverty and inequality, demography, income and consumption expenditure, education, employment, health, housing and durable goods, participation in poverty alleviation programs. These welfare indicators are estimated for the whole country, separately for urban and rural areas, and for eight socioeconomic regions. The regions consist of: Red River Delta, North East, North West, North Central Coast, South Central Coast, Central Highlands, North East South, Mekong River Delta.

The reports have the length from 30 to 50 pages, and made in printout and electronic formats (CD-ROM or diskette) depending on the users' requirement.

- (3) Completed reports of VHLSS results: These reports provide detailed information on socioeconomic characteristics of household and community. In addition to general indicators of socioeconomic issues as given by the preliminary reports, the completed reports presents the detailed analysis of each issues. For example, while the preliminary reports provide only estimates of income level, the completed reports break down income by income sources, household size, gender and age of household head. The results are estimated at the regional and provincial level.

The reports have the length from 80 to 120 pages, and made in printout and electronic formats (CD-ROM or diskette) depending on the users' requirement.

- (4) A book of VHLSS results: After receiving comments on the completed reports, the GSO will revise and add more detailed results of statistical analysis of living standards using VHLSS data to produce a book "Results of Viet Nam Household Living Standards Surveys". The book contains from 10 to 15 chapters with a length of around 500 pages, providing insight into the monitoring and evaluation of living standards, trends in poverty and inequality using data tables, illustrative figures and charts. It is printed in both Vietnamese and English, and published by the Statistical Publishing House.
- (5) Required statistical products: Being aware that the micro-data is difficult to get used to, the GSO will provide results of the statistical analysis as required by particular users. The users can request detailed analytical reports based on VHLSS data including cross-tabulation tables, figures, charts, and statistical interpretation. The required reports are made in the form, printout or electronic (CD-ROM or diskette), depending on the users' need.

In order to increase the effectiveness of the data use the GSO will conduct presentations, training courses related to the use of the VHLSS as required by the users. Technical assistance in analyzing the VHLSS data is also provided to help the users reduce their time in getting used to the data and increase the efficiency and science of their research. However, if the users are familiar with the format of the data set, they can obtain the data set and use it for their purposes without any requirement related to training assistant.

- (6) Micro-data set: After deleting the identification variables to protect the confidentiality the micro-data set of VHLSS including the household and commune data will be provided for the users. In general the data are grouped into different files by the thematic sections in the questionnaire, e.g. education or employment. Identity codes for respondents including household members, households, and communes are generated to merge the data between data files. The codes can be used to link the data from different VHLSS to create the panel data. The whole data set or a portion of it can be provided for the users depending on their need. The GSO also prepares aggregated files with variables selected by the users.

The micro-data of VHLSS are stored using program Stata. The Stata files can be easily transferred into other file formats such as SPSS, EXCEL, SAS, etc using data-transfer programs.

#### Dissemination Calendar

The GSO makes every effort to disseminate the survey documents, reports and micro-data of VHLSS as soon as possible following data collection, subject only to limits imposed by resources, technology, and data quality. The GSO will not restrict the dissemination in order to preserve publication rights of its staff, donor agencies, or other State organizations. Prior to the dissemination of VHLSS the GSO will thoroughly evaluate data quality and assure that the data release will preserve the respondents' confidentiality. Expert assistance is often needed to conduct data quality reviews. Procedures for conducting such reviews should be a part of all data planning activities.

For basic documents of VHLSS: The basic documents on the survey design and data collection such as questionnaire, sampling methods, interview manuals will be provided for the user after the data collection is completed. Other documents related to the data processing and cleaning are disseminated as soon as they are produced completely.

For result reports: VHLSS is a large survey collecting a huge amount of information. It is estimated that the preliminary processing and cleaning of data might take from four to six months. The statistical reports using VHLSS data on topics required by particular users are expected to be provided for the users four months after the completion of data collection. The preliminary reports will be disseminated six months after the completion of data collection. Then preliminary reports will be revised and added more detailed information on living standards to produce completed reports. These reports are released nine months after the data collection is completed.

The publication of the book “Results of Viet Nam Household Living Standards Surveys” requires more time for data analysis, editing and printing. The book is planned to be published one and half year after the completion of data collection.

For micro-data: The GSO strives for making the VHLSS micro-data set available as promptly as possible, subject only to limits imposed by resources, technology, and data quality. As a State agency, the GSO has to be fully responsible for the disseminated data. They are required to ensure the quality of VHLSS data and protection of respondent. For the living standard survey, the requirement of micro-data equality is utmost because the analysis of micro-data results in socioeconomic implications which might have strong effects on the living standards of the people. The VHLSS micro-data is reviewed carefully by the GSO and assistance agencies by comparing the results extracted from these data with other information sources. When there are suspicions about the accuracy of data, the GSO will have to review the data from the collection to the computer entry. If the data cannot be corrected they have to be deleted, and the weight might be adjusted to ensure the representativeness of the data set. The evaluation of data quality is conducted during the processing and analyzing of the data. This substantial work can take a long time. The micro-data of VHLSS are proposed to be released to the users two years after the data collection is completed.

There are, however, situations where it would be beneficial to release a portion of the micro-data or aggregated data prior to the time when the full set of micro-data can be made available. The requirement for such “early releases” or a staggered release can be raised by the GSO and donor agencies to fulfill important policy and scientific goals.

#### *6.2.2. Recipients and Dissemination Mechanism*

The GSO aims to disseminate the data and information of VHLSS as widely as possible, subject only to limits imposed by resources, technology, data quality, and confidentiality protection. All sorts of people are ensured the equitable accessibility to the reports and micro-data of VHLSS.



For the document and analytical reports of VHLSS: The survey documents and reports of VHLSS results including preliminarily and completed reports as specified above are made available to the public. However in the context of limited human and material resources the GSO will require the users to meet provision conditions to assure that the users do not misuse the data and not ask for more data than they are likely to use. The conditions are applied as follows:

- For State organizations, international agencies with offices in Vietnam, domestic institutes and universities and their staffs, researchers and students, the documents and reports are made readily if they hand in a formal letter of data request. This letter should State clearly the purpose of data use. A small fee will be charged to cover the costs of printing and mailing.
- For other users, in addition to the fee of printing and mailing a modest fee of data distribution will be charged depending on the required amount of documents and reports in order to prevent the data misuse that could affect the human and material resources of the GSO. The fee does not include the cost of data collection, computer entry, data processing and cleaning, and data analysis.

The book “Results of Viet Nam Household Living Standards Surveys” is published for sale.

Required statistical products: The GSO provides particular analytical reports, statistical assistance, presentations and training courses according to users’ requirements. A service fee to cover the cost of data analysis, report writing, and assistance will be charged to the users. The amount of fee will depend on the specific requirements of the users.

Micro-data set: The micro-data set of VHLSS will be disseminated to the users on the payments basic with two access policies: data use agreements and the controlled access to micro-data:

- Data Use Agreements: The micro-data are only provided for the users if they commit themselves to data use agreements. No data which cannot be publicly released will be made available without a data user’s signed written agreement to provide such safeguards as are necessary. In general the use data agreements include points: (i) The users are prohibited to modify and change the micro-data. The research reports using VHLSS data are required to specify the data source from VHLSS. (ii) The users are not allowed to use the micro-data to track the identification of the respondents. (iii) The users are required not to pass the data set to third parties for any reason. (iv) If the users obtain the data set without payment, they have to hand in the proposal of using the data, and commit themselves to using the data only for the specified purposes. Further use of the micro-data for new researches will need the approval of the GSO. A copy of research reports using the VHLSS data is required to be sent to the GSO.
- Controlled access: If the users conduct research assignments for the GSO and donor agencies that rely on the VHLSS micro-data, users will have access to the data without any payment but are not in possession of the data. The GSO and donor agencies exercise direct supervision of the data use in order to protect the misuse of data. When the assignments are completed the GSO might ask the users to return the data set.

For the State organizations, the micro data will be provided through the mechanism of data use agreements two years after the data collection. A small amount of fee will be charged to cover the cost of producing CD/diskette and delivery.

For other users, the full micro-data set will be supplied on the payment basis. The users are required not to track the respondents identification, to modify the micro-data, and to pass the data to third parties. The data processing and cleaning require substantial resources. Thus the charge aims to cover a fraction of costs of sorting and cleaning information, and distributing the data set. In addition the charge also helps to ensure that the use of the micro-data is necessary and the potential users have an incentive not to ask for more data than they are likely to use. The cost of data collection is not included in the charges.

Contact:

Mr. Nguyen Phong  
Director  
Social & Environment Department  
General Statistical Office  
No 2- Hoang Van Thu- Hanoi  
Phone: 8433354  
Email: [nphong@gso.gov.vn](mailto:nphong@gso.gov.vn)  
cc: [ntquan@gso.gov.vn](mailto:ntquan@gso.gov.vn)