



YOUNG LIVES TECHNICAL NOTE NO. 4
March 2008

An Assessment of the Young Lives Sampling Approach in Vietnam



Ngoc P. Nguyen

Contents

Executive summary	2
1. Introduction	4
2. Sampling frame in Vietnam	4
2.1 Selection of provinces	5
2.2 Selection of sentinel sites	6
3. Representativeness of Vietnam Young Lives first round data	7
3.1 Comparison of Young Lives with the Vietnam Household Living Standard Survey	7
3.2 Comparison of Young Lives with the Demographic and Health Survey	10
4. How relatively poor is the Young Lives sample?	12
5. Non-parametric tests for equality of two distributions	14
6. Conclusion	16
References	17
Appendix 1: Young Lives study sites in Vietnam	18
Appendix 2: Non-parametric tests for equality of two distributions	19
Appendix 3: Calculation of household wealth index	20
Appendix 4: Comparison of Young Lives 1-year-old sample with the Vietnam Household Living Standard Survey 2002	21
Appendix 5: Comparison of Young Lives 8-year-old sample with the Vietnam Household Living Standard Survey 2002	23
Appendix 6: Comparison of Young Lives with the Demographic and Health Survey 2002	25

Executive summary

Young Lives is a longitudinal research project investigating the changing nature of childhood poverty. The study is tracking the development of 12,000 children in Ethiopia, Peru, India (Andhra Pradesh) and Vietnam through qualitative and quantitative research over a 15-year period. Since 2002, the study has been following two cohorts in each study country. The younger cohort consists of 2,000 children per study country aged between 6 and 18 months in 2002. The older cohort consists of 1,000 children per country aged between 7.5 and 8.5 in 2002. The key objectives of Young Lives are: (i) to improve the understanding of causes and consequences of childhood poverty, (ii) to inform the development and implementation of future policies and practices that will reduce childhood poverty.

The sampling approach adopted for the Young Lives study is known as a sentinel site surveillance system. In Vietnam the Young Lives team used multi-stage, purposive and random sampling to select the two cohorts of children. This methodology randomised households within a study site while the sites themselves were chosen based on predetermined criteria, informed by the objectives of the study. To ensure the sustainability of the study, and for resurveying purposes, a number of well-defined sites were chosen. The selection of sentinel sites in Vietnam followed a complex consultative process in which five 'representative' regions were selected on the basis on the regional and urban/rural differences present in modern Vietnam. Next, four communes from each region were selected using a pro-poor selection rule based on the poverty ranking of each commune.

This paper assesses the sampling methodology by comparing the Young Lives sample with larger, nationally representative samples. In doing this, the Vietnam team sought to:

- analyse how the Young Lives children and households compare with other children in Ethiopia in terms of their living standards and other characteristics
- examine whether this may affect inferences between the data
- establish to what extent the Young Lives sample is a relatively poorer or richer sub-population in Vietnam
- determine whether different levels of living standards are represented within the dataset.

We used two nationally representative comparison samples, the Vietnam Household Living Standard Survey 2002 (VLSS 2002) and the Demographic and Health Survey 2002 (DHS 2002), and applied two different methodologies to assess the Young Lives sample. We first compared wealth index scores for the Young Lives households with those for DHS households. This provided a graphical illustration of the relative wealth of the Young Lives sample relative to the population of Vietnam. We went on to use standard t-tests to test for statistical significance of the differences in several living standard indicators between Young Lives, the DHS 2002, and the VLSS 2002 samples. In order to ensure comparability of the different samples we imposed constraints on the comparison samples to accommodate first the fact that the Young Lives sample only includes households with at least one child aged between 6 and 18 months or aged 8 years, and second that the site selection methodology means that the regional and rural/urban weights implied by the Young Lives sample will not reflect the true population weights.

We found that households in the Young Lives sample were poorer than households in the DHS 2002 sample. A similar picture emerged when we use t-tests to compare the means for a range of living standard indicators between the Young Lives, the VLSS 2002 and the DHS 2002 samples. Young Lives households own less assets, are less likely to own their own

house, and have access to fewer services and utilities than the VLSS households. They are also more likely to be registered as poor by their local authorities, and be classified as 'hungry and poor'. However, household heads in the Young Lives sample were more likely to be male and more educated than the national average, while the primary caregiver were more likely to be female and less educated. We conclude that the results reflect the pro-poor sampling approach chosen by Young Lives in Vietnam.

In summary, consistent with the sampling methodology applied in Vietnam we find that Young Lives households are generally poorer than the average Vietnamese household. Despite these biases, it is shown that the Young Lives sample in Vietnam covers the diversity of children in the country in a wide variety of attributes and experiences. Therefore, while not suited for simple monitoring of child outcome indicators, the Young Lives sample will be an appropriate and valuable instrument for analysing causal relations, modelling child welfare, and its longitudinal dynamics in Vietnam.

1. Introduction

Young Lives is a longitudinal research project investigating the changing nature of childhood poverty. The study is tracking the development of 12,000 children in Ethiopia, Peru, India (Andhra Pradesh) and Vietnam through qualitative and quantitative research over a 15-year period. Since 2002, the study is following two cohorts in each study country. The younger cohort or 1-year-old cohort consists of 2,000 children per study country aged between 6 and 18 months in 2002. The older cohort or 8-year-old cohort consists of 1,000 children per country aged between 7.5 and 8.5 years in 2002. The key objectives of Young Lives are: (1) to improve the understanding of causes and consequences of childhood poverty, (2) to inform the development and implementation of future policies and practices that will reduce childhood poverty.

To fit the main objectives of the project, Young Lives employed a sentinel site sampling method, which is a multistage sampling approach and uses both purposive and probability sampling methods. While households within the sites were selected by a method equivalent to random sampling, the site selection process was not random and over sampling of poor sites took place.

The aim of this report is to assess the chosen sampling methodology. Therefore, the report investigates the representativeness of Young Lives data by comparing selected variables from the Young Lives sample with those from two nationally representative surveys: the Vietnam Household Living Standard Survey 2002 (VLSS 2002) and the Demographic and Health Survey 2002 (DHS 2002). The aim of the comparison is to establish to what extent the Young Lives sample in Vietnam is a poorer or richer sub-population.

The structure of this report is as follows. Section 2 describes the sampling design adopted by Young Lives Vietnam. Section 3 discusses the representativeness of Young Lives' first round data in Vietnam. Section 4 and 5 compare the Young Lives sample with the DHS sample using different methodologies including comparison of distribution of wealth index by household. Section 6 concludes.

2. Sampling frame in Vietnam

In September 2001, the research team in Vietnam selected the Young Lives study sites using a sentinel site sampling approach. In Vietnam, a sentinel site was defined as a commune. A commune has a local government, primary school, commune health centre, post office, and a market. In the event that a commune selected as a sentinel site had insufficient numbers of 1-year-old children at the time of the survey, a neighbouring commune with similar socioeconomic conditions was additionally selected in order to reach the quota of children. Therefore, with 20 sentinel sites, Vietnam had 31 communes involved in the study sample (Tuan et al. 2003).

Vietnam followed a five-step process in its sampling strategy:

1. Selection of five regions out of nine regions in Vietnam.
2. Selection of one province out of all provinces in each chosen region.
3. Selection of four sentinel sites in each province with oversampling of poor sites.

4. Screening of all households in each sentinel site and listing of eligible children. Only single children were included and families with twins and triplets were excluded. Where there was more than one eligible child present in the household, one of those children was selected using a random sampling technique (Wilson et al. 2006).
5. Selection of a sample of 100 1-year-old and 50 8-year-old children in each sentinel site using simple random sampling.

2.1 Selection of provinces

In 2001, Vietnam had 61 provinces and cities, which were divided into 600 districts and 10,321 communes.¹ In terms of socioeconomic development, Vietnam can be stratified into eight socioeconomic regions: North-West, North-East, Red River Delta, North Central Coast, South Central Coast, South-East, Central Highlands, and Mekong River Delta. Additionally, the Young Lives team in Vietnam categorised all major urban centres (Hanoi, Ho Chi Minh City, Da Nang, Hai Phong, and Ba Ria-Vung Tau) as a new region – the Cities region (Tuan et al. 2003).

Box 1: *Criteria for selecting regions*

1. Consist of regions in the North, Central and South
2. Consist of urban, rural, and mountainous area
3. Be over-poor
4. Reflect some unique factors of the country, such as natural disaster and war consequences

Out of these nine regions, five regions were selected through a process of iterative consultation with many different parties² using various methods such as individual interviews and group meetings. The criteria for selection are presented in Box 1. From each region, one province was selected. The provinces selected were Lao Cai (North-East region), Hung Yen (Red River Delta), Da Nang (Cities), Phu Yen (South Central Coast), and Ben Tre (Mekong River Delta) (see Appendix 1).

Table 1. *Socioeconomic indicators of the five provinces selected for Young*

Province	Human Development Index rank ^a	Total population (1999)	Migration (1994-99)	Population under poverty line ^b (%)	GDP per Capita (US\$)	Infant mortality rate (% ,1999)	Underweight children under age 5 (% ,1998)	Population without access to safe water (% ,1999)	Population without access to sanitation (% ,1999)	Ethnic minority (% ,1999)
Lao Cai	55	594,632	-415	22	144	5.36	46	54	43	67
Hung Yen	14	1,068,704	-14,886	13	209	2.56	40	6	3	0.1
Da Nang	4	684,131	24,692	12	409	1.90	38	5	15	1
Phu Yen	49	786,969	-3,659	9	202	4.30	41	10	12	10
Ben Tre	27	1,296,914	-34,816	22	258	4.08	34	80	3	0.4

Notes: ^a Ranking out of 61 provinces and cities with descending order (1=best).

^b The national poverty line in Vietnam was 12.9 per cent in 2002.

Source: NCSSH 2001

¹ The most up-to-date figure for Vietnam is 64 provinces and cities (Vietnam Government Website 2007).

² Including Principal Investigator (PI), National Coordinator (NC), Technical Advisory Committee (TAC) and Policy Advisory Committee (PAC) members who represent 29 government institutions, international donors, and local NGOs.

Table 1 summarises socioeconomic indicators of the five selected provinces. Based on the presented indicators Da Nang can be classified as above average, followed by Hung Yen and Phu Yen (average) and lastly, Lao Cai and Ben Tre (below average). In terms of Human Development Index (HDI),³ Phu Yen and Lao Cai have the lowest HDI at rank 49 and 55 out of 61 provinces and cities, respectively. Hung Yen and Ben Tre are in the middle of the ranking (14th and 27th) while Da Nang, as a city, has a HDI near the top (4th).

In Lao Cai, Hung Yen, and Ben Tre more than the national average of 12.9 per cent of the population lived under the poverty line.⁴ Lao Cai stands out as the province with the highest proportion of ethnic minority population, 67 per cent compared with the national average of 16 per cent, and tops the list in terms of population without access to sanitation, proportion of underweight children under five, and infant mortality. In Ben Tre 80 per cent of the population did not have access to safe water.

2.2 Selection of sentinel sites

The selection of four sentinel sites in each province was carried out by the provincial governments. For this purpose, provincial working groups were established. The groups ranked all communes in the province by poverty level: poor, average, better off and rich. Criteria used for ranking included: (1) development of infrastructure, (2) percentage of the poor households in the commune, (3) child malnutrition status.

Four sentinel sites were selected using an over-poor sampling strategy: two communes from the poor group, one from the average, and one from the above average group (combined better off and rich). Other criteria used in the selection are presented in Box 2. Among the 31 communes selected, 15 communes were from the poor group (48 per cent), nine communes were from the average (29 per cent), and seven communes (23 per cent) were from the above average.

A total of 25 out of 31 communes (81 per cent) had a cement or tar road to the centre of the commune. All communes had a commune health centre and elementary school. There were three communes in mountainous and remote areas, where access to a hospital was described as difficult (Tuan et al. 2003).

Box 2: *Criteria for selecting sentinel sites*

1. Over-poor sampling strategy
2. Represent common provincial/regional features
3. Commitment from the local government for the research
4. Feasibility conditions for the research logistics
5. Population size

³ Human Development Index (HDI) is a comparative measure of life expectancy, literacy, education and standard of living. In Vietnam a country specific HDI was calculated for each province.

⁴ In 2002, the national poverty line in Vietnam was defined as 150,000 Vietnam Dong (VND) per month for urban areas, 100,000 VND per month for rural areas, and 80,000 VND for mountainous and remote regions (NCSSH 2001).

3. Representativeness of Vietnam Young Lives first round data

Due to the non-random sampling procedure, significant bias in the Young Lives sample can be expected. Furthermore, the modest size of the samples (2000 1-year-olds and 1000 8-year-olds) and the purposively oversampling of poor sites may further influence the representativeness of the sample.

Wilson et al. (2006) acknowledge that Young Lives has limitation as a source for national level summative information. They argue that Young Lives is intended much more as an in-depth study of relationships between pieces of information, rather than an instrument to collect national statistical results, as is the requirement from the more traditional systems.

In a simple random sampling framework, sampling units like communities or sub-districts are not interesting in themselves, but are just part of the structure used to sustain claims of randomness of sampling and hence to generalisation of summative quantitative measures to the population, strata or zones. In the sample framework adopted by Young Lives each sentinel sites is interesting in their own right and can provide views of particular aspects.

A general population sample would not give sufficient attention to the target group of poor children. Therefore, over-sampling of poor sites took place. However, not only poor sites but also non-poor sites were chosen to enable comparison (Wilson et al. 2006). The Young Lives sampling method was not designed to equally represent the whole national population of children born in the qualifying period but to reflect the purpose of the study, and yet be logical and statistically viable.

The objective of this report is to validate the representativeness of certain features of the sample by comparing Young Lives data with data from two nationally representative samples. There are no other surveys in Vietnam that have focused on a sub-population of young children like the Young Lives survey.

The nationally representative samples are from the Vietnam Household Living Standard Survey 2002 (VLSS) (GSO 2003) and the Vietnam Demographic and Health Survey 2002 (DHS) (CPFCV 2003). Both surveys were carried out during the same time as the first round data collection of Young Lives and followed a random multistage stratification strategy. The sample size of both surveys is huge. The VLSS sample includes nearly 30,000 households and covers all 61 provinces and cities in Vietnam, the DHS sample contains more than 7,000 households in all regions. The VLSS 2002 provides detailed information on household composition and characteristics whereas the DHS 2002 has information on child and maternal care.

3.1 Comparison of Young Lives with the Vietnam Household Living Standard Survey

In the comparison exercise, 17 common variables from each sample were chosen. The variables can be divided into three categories: household characteristics, head of household characteristics and individual characteristics of household members (see Box 3). These indicators will allow us to compare household composition, socioeconomic status and individual characteristics of the Young Lives samples with the VLSS samples.

The Young Lives 1-year-old and 8-year-old cohorts are compared with the VLSS sample at both national and regional level. In order to compare each cohort with the same specific population group, the national and regional samples are restricted to include only households with 1-year-old or 8-year-old children.

(i) Young Lives 1-year-old cohort

Appendix 4 presents the results for all comparison tests carried out between the Young Lives sample and (i) the VLSS national sample, (ii) the national sample of households with 1-year-old child, (iii) the regional sample,⁵ and (iv) the regional sample of households with 1-year-old child. Some trends emerged which will be discussed in the following.

Household characteristics

Households from the Young lives sample were almost three times more likely to be categorised as poor or hungry than households from the VLSS sample. Moreover, 7 per cent of the Young Lives households were certified as poor household by local authorities, compared with 4.14 per cent of regional VLSS sample of households with 1-year-old child.

Box 3: *Comparison variables, Young Lives and VLSS 2002*

HOUSEHOLD CHARACTERISTICS

1. Household size
2. House ownership
3. Electricity supply
4. Source of drinking water
5. Sanitation facility
6. Radio ownership
7. TV ownership
8. Household category
9. Certified as poor household by local authorities
10. Livestock ownership in the past twelve months

HEAD OF HOUSEHOLD CHARACTERISTICS

1. Age
2. Sex
3. Education
4. Marital status

INDIVIDUAL CHARACTERISTICS

1. Age
2. Sex
3. Education

Households in the Young Lives sample were six times more likely not to own a house compared with the national average and four times more likely compared with the regional sample.

Households in the Young Lives cohort were less likely to have access to improved sanitation facilities and drinking water sources compared with the national average. For example, nearly 41 per cent of Young Lives households used unprotected sources of drinking water such as water from springs or ponds compared with 18 per cent in the VLSS national sample and 22 per cent in the VLSS regional sample with 1-year-olds. A household in the Young Lives

sample consisted of an average of 4.9 members, slightly higher than the national and regional average. However, it is lower than the comparable groups of households with 1-year-old child, both nationally and regionally. The difference in household size was relatively small.

Head of household characteristics

The household head in the Young Lives sample was 20 per cent younger than the national average. The reason might be that Young Lives focuses on young families with small children.

There were more male-headed households in the Young Lives sample than in the VLSS samples. On average, 84 per cent of households in Young Lives were headed by a man and only 16 per cent were headed by a woman.

Since Young Lives focuses on young families, a higher proportion of household heads was married and significantly, less household heads were widowed compared with the VLSS samples.

Nearly 70 per cent of household heads had some schooling and qualification compared with the national average of 60 per cent. Twice as many household heads in the Young Lives sample had a university degree or higher education as in the VLSS national sample.

Individual characteristics of household members

The average age of a household member in the Young Lives sample was slightly higher than the national average in VLSS sample. Out of nearly 7,300 individuals in the Young Lives sample, about 53 per cent were female and 47 per cent male. Comparing to the VLSS samples, the benchmark for the male-female sex ratio was 50-50.

On average household members in Young Lives project seemed to have more education than those in the VLSS sample. The proportion of household members without schooling or qualification was 10 per cent less in the Young Live sample compared with the VLSS sample. Around 57 per cent of individuals in the Young Lives sample achieved primary and secondary education, compared with 44 per cent in the VLSS regional sample with 1-year-old child.

(ii) Young Lives 8-year-old cohort

Appendix 5 displays the results for all comparison tests carried out between the Young Lives 8-year-old sample and (i) the VLSS national sample, (ii) the national sample of households with 8-year-old child, (iii) the regional sample⁶, and (iv) the regional sample of households with 8-year-old child. There is a remarkable similarity in characteristics of the Young Lives 8-year-old sample and the 1-year-old sample.

Household characteristics

Fewer households in Young Lives 8-year-old cohort owned a house than the national average. Furthermore, a significantly higher proportion of Young Lives households used unprotected sources of drinking water and had no toilet facility. One-third of the Young Lives sample was categorised as hungry and poor household compared with one-tenth of VLSS regional sample of households with 8-year-olds. Authorities certified 10 per cent of Young Lives households with 8-year-old as poor households compared with only 3.7 per cent in the VLSS national sample.

⁶ The same five provinces as the Young Lives provinces.

Head of household characteristics

The characteristics of household head were very similar to those in the Young Lives 1-year-old sample. The average age of a household head was lower than the national average, a higher proportion of households was male-headed, the head was on average more educated and a higher proportion was married while less were widowed.

Individual characteristics of household members

The characteristics of household members in the 8-year-old cohort were similar to those in the 1-year-old cohort. The average age was 26.7 years, which was lower than in the national and regional sample but higher than in the VLSS samples of households with 8-year-old child. The male-female sex ratio was 48-52, which was similar to the national average and very similar to the VLSS regional sample with 8-year-olds. On average household members in the Young Lives 8-year-old cohort were also more educated than those in the VLSS sample. Twice as many household members in the VLSS regional sample with 8-year-olds had no schooling or qualification as in the Young Lives sample.

3.2 Comparison of Young Lives with the Demographic and Health Survey

Box 4: *Comparison variables, Young Lives and DHS 2002*

HOUSEHOLD CHARACTERISTICS

1. Average household size
2. Electricity supply
3. Source of drinking water
4. Sanitation facility
5. Radio ownership
6. TV ownership

HEAD OF HOUSEHOLD CHARACTERISTICS

1. Age
2. Sex
3. Education
4. Marital status

PRIMARY CAREGIVER BACKGROUND

1. Literacy rate
2. Highest grade completed in school
3. Ethnic background
4. Religion

PREGNANCY, DELIVERY AND BREASTFEEDING

1. Received antenatal care
2. Number of antenatal visits during pregnancy
3. Received two or more Tetanus injections during pregnancy
4. Place of delivery
5. Breastfeeding

The DHS enables us to compare characteristics of primary caregivers and on pregnancy, delivery, and breastfeeding with the Young Lives sample. This information was not collected in the VLSS. Besides, the DHS provides additional information on household characteristics and head of household characteristics. A total of 19 variables under four categories was chosen for comparison (Box 4).

Appendix 6 shows the summary statistics of these 19 variables across the Young Lives 1-year-old sample, the 8-year-old sample and the DHS 2002 sample. The sample size for the household characteristics and head of household characteristics was 2,000 households for Young Lives 1-year-old cohort, 1,000 households for the 8-year-old cohort and 7,048 households for the DHS sample. For the other two categories of comparison variables the sample size in DHS was reduced to 5,665 households that is equivalent to the number of ever-married women aged 15-49 interviewed in the survey.

Household characteristics

On average, a Young Lives household consisted of 4.9 members, slightly higher compared with 4.4 in the DHS sample. Similar proportions of households in the Young Lives and the DHS sample had electricity supply and owned a radio. There was no variable in the DHS on the number of households classified as poor. Nevertheless, the higher percentage of households that used unprotected sources of drinking water, did not have sanitation facilities, and did not own a TV in the Young Lives samples compared with households in the DHS implies that there might have been more poor households in the Young Lives sample.

Head of household characteristics

The heads of household in Young Lives sample were on average younger, more educated, more likely to be married than those in the DHS sample. A total of 84 per cent of Young Lives households were headed by a man compared with 72 per cent in the DHS. The data clearly shows that Young Lives focuses on young families with small children.

Primary caregiver background

The backgrounds of primary caregivers greatly influence the way a child is raised. Around 13 per cent of primary caregivers in the Young Lives sample were non-literate, compared with 8.5 per cent in the DHS sample. Household heads in the Young Lives sample appear to have more education than the DHS counterparts have. Given that, 97 per cent of primary caregivers in Young Lives were the child's biological mothers, this shows that the male population in the five Young Live provinces was on average more educated than the female population.

The data on ethnic background and religion are very similar across the three samples. The majority of primary caregiver in Young Lives and DHS samples are of Kinh ethnicity and have no religion. About 16 per cent of caregiver in Young Lives 1-year-old sample belongs to ethnic minority groups, the same as in the DHS sample.

Pregnancy, delivery and breastfeeding

Since the Young Lives 8-year-old sample was not asked questions on this section, it is only possible to compare the Young Lives 1-year-old sample with the DHS. A total of 82 per cent of mothers in the Young Lives study received antenatal care, slightly lower than the 86 per cent in the DHS sample. This trend is also shown by a higher proportion of Young Lives mothers with no antenatal visits during pregnancy, 18.25 per cent compared with 13.2 per cent in the DHS. Fewer mothers in Young Lives sample received two or more injection for Tetanus during pregnancy compared with mothers from the DHS sample.

Breastfeeding and place of delivery were very similar in the Young Lives and the DHS sample. Almost every woman (98 per cent) breastfed and every second woman gave birth at hospital and about 20 per cent delivered at home. The remaining 30 percentage gave birth in commune health facilities.

4. How relatively poor is the Young Lives sample?

The comparison of the Young Lives sample with the DHS and the VLSS samples points to a systematically poorer sample of households in the Young Lives survey. To see in more details how poor the Young Lives sample is relative to the national statistics, we plot the cumulative distribution functions (cdf) of the household wealth index⁷ for the Young Lives 1-year-old sample and the DHS sample. As can be seen in Figure 1, all the points on the Young Lives cdf lie at least as high or above those in the DHS cdf for any given value of wealth index. This shows a first order dominance, meaning the YOUNG LIVES sample is poorer than the DHS sample.

Figure 1. Cumulative distribution functions of household wealth index, Young Lives and DHS 2002, household-to-household

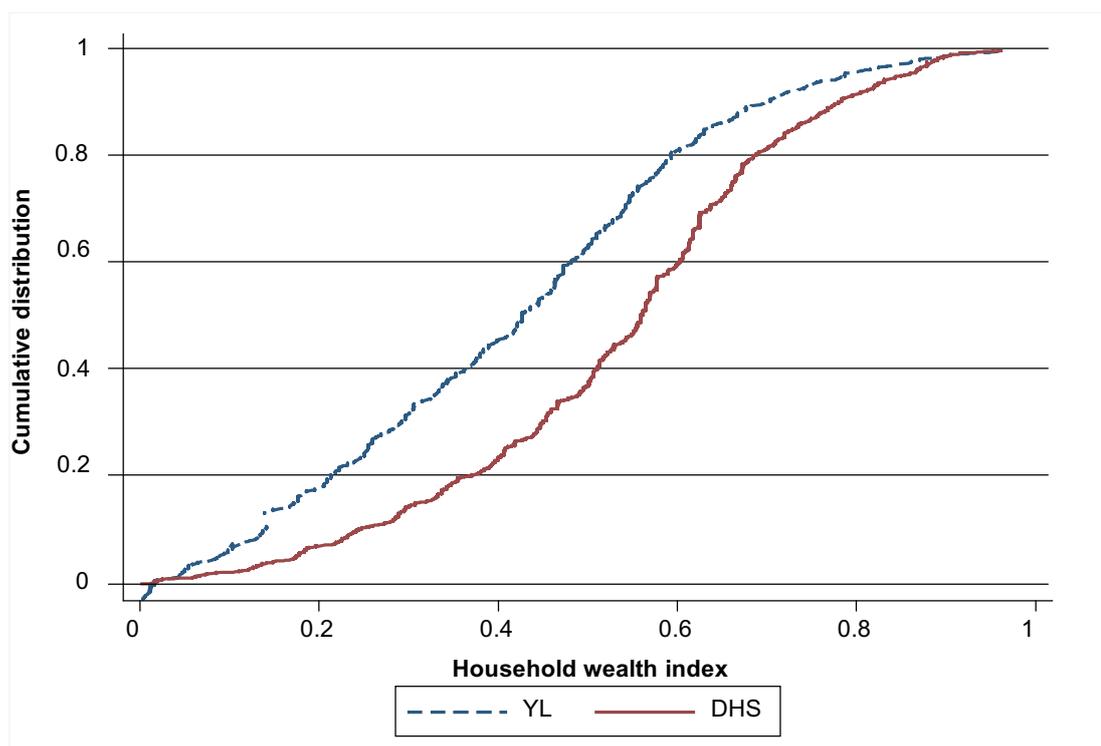
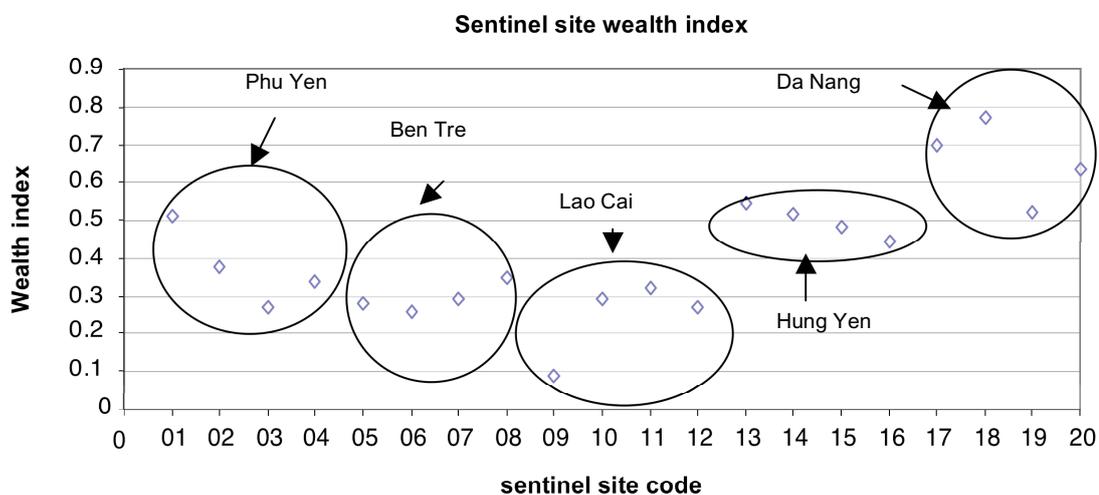
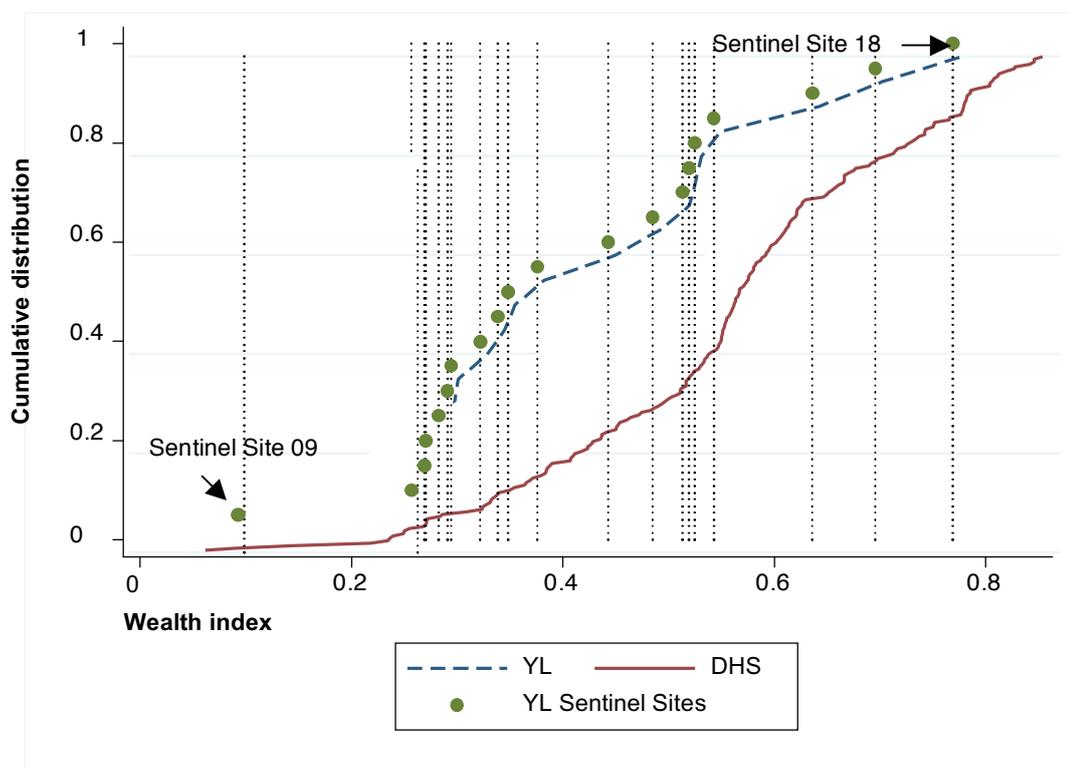


Figure 2. Wealth index distribution by sentinel site and province**Table 2.** Ranking of Young Lives sentinel sites by wealth index

Ranking	Site Code	Province	Site Wealth Index	Below the mean
1	09	Lao Cai	0.092584	Yes
2	06	Ben Tre	0.256875	Yes
3	03	Phu Yen	0.268855	Yes
4	12	Lao Cai	0.270362	Yes
5	05	Ben Tre	0.282637	Yes
6	07	Ben Tre	0.290869	Yes
7	10	Lao Cai	0.294414	Yes
8	11	Lao Cai	0.322011	Yes
9	04	Phu Yen	0.338579	Yes
10	08	Ben Tre	0.348203	Yes
11	02	Phu Yen	0.376087	Yes
12	16	Hung Yen	0.443091	Yes
13	15	Hung Yen	0.484868	Yes
14	01	Phu Yen	0.513294	No
15	14	Hung Yen	0.519408	No
16	19	Da Nang	0.524648	No
17	13	Hung Yen	0.543116	No
18	20	Da Nang	0.636078	No
19	17	Da Nang	0.696063	No
20	18	Da Nang	0.768902	No

In Figure 3, the 20 Young Lives sentinel sites are ranked according to wealth index scores and placed along the cumulative distribution functions for comparison. The poorest sentinel site in the sample is site 09 in Lao Cai province. Its wealth index score lies in the bottom 10 per cent of the Young Lives distribution and in the bottom 5 per cent of the DHS distribution. More than 60 per cent of the Young Lives sentinel sites have a wealth index score below the mean compared with only 30 per cent of the DHS sample. In conclusion, Young Lives sentinel sites are systematically poorer than the DHS sample.

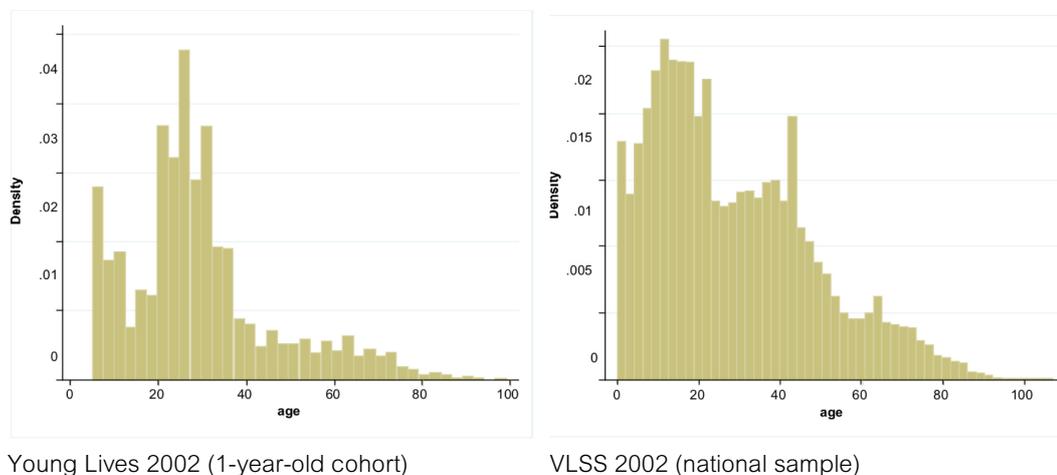
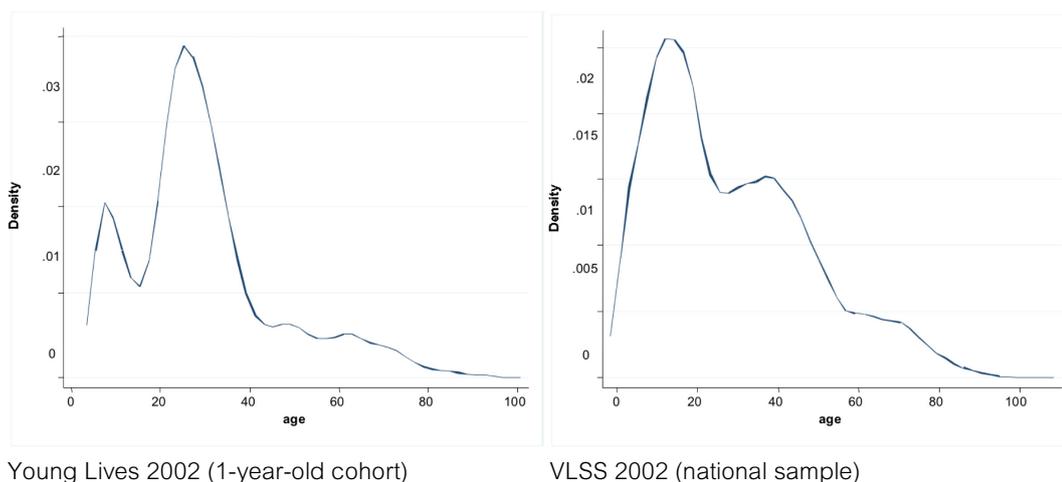
Figure 3. Cumulative distribution functions of household wealth index by sentinel site, Young Lives and DHS 2002



5. Non-parametric tests for equality of two distributions

The summary statistics are useful in sketching a general picture of household and individual characteristics, but they cannot determine whether the distribution of a variable is equal to the distribution in another sample. For example, the individual age which was collected in the Young Lives sample and the VLSS sample. There were 7,294 observations in the Young Lives 1-year-old sample with a mean age of 29.54 years. The VLSS national sample included 132,384 individuals with a mean age of 28.57 years. If we just compare the means between the two samples by t-test, the difference is small and insignificant. However, this is not sufficient to determine the difference in distributions of the variable across two independent samples. The differences in size and distribution need to be taken into account. The histogram and Epanechnikov Kernel density estimates⁸ for the age variable reveal a very different picture in the distributions of age between the two samples (Figure 4 and 5).

⁸ A Epanechnikov Kernel density estimate is a statistical method widely used to estimate density functions. A density function is similar to a histogram for continuous data. For the kernel estimation, a smoothing parameter called the bandwidth (h) is used. In STATA, the default value for h is 0.545 (Epanechnikov bandwidth) (Cameron and Trivedi 2005).

Figure 4. Age distribution of household members, Young Lives and VLSS 2002**Figure 5.** Epanechnikov Kernel density estimate of age, Young Lives and VLSS 2002

In order to test whether the two data sets are drawn from the same population distribution or from different distribution functions, we use three non-parametric tests: (1) two-sample Kolmogorov-Smirnov test of the equality of distributions, (2) Wilcoxon ranksum or Wilcoxon-Mann-Whitney test for the same distribution, and (3) Median chi-square test on the equality of medians.⁹ The advantage of using non-parametric tests is that they do not require any assumptions about the distribution of data.

The null hypothesis is that two independent samples are drawn from populations with the same distribution or the same medians. *Reject* the null hypothesis means that the two datasets are drawn from two different distributions or that the medians are not identical. Failure to reject (*cannot reject*) the null hypothesis shows that the two datasets are consistent with a single distribution function or the medians are the same.

Since there is a large degree of consistency between the three tests they tend to give the same results for each variable (either reject or cannot reject), only the results of the Kolmogorov-Smirnov test for the equality of two distributions are presented in Appendix 4.

⁹ See Appendix 2 for more details on each test.

Distributions of each variable are shown to be statistically different from each other at 1 per cent, 5 per cent or 10 per cent significant level.

A trend can be noticed. The number of significantly different variables seems to be reduced when the comparative sample is broken down further into regional sample or samples of households with 1 or 8-year-old child. For example, when we compare household characteristics across the Young Lives 1-year-old sample and VLSS 2002 samples, the number of statistically different variables goes down from 17 in the national sample to 16 in the national sample of household with 1-year-old child, 14 in regional sample and finally, 11 in the regional sample of households with 1-year-old child.

6. Conclusion

The purpose of this report is to compare Young Lives first round data with statistics from nationally representative samples to establish the representativeness of Young Lives sample in Vietnam. We have done this by both comparing summary statistics and performing non-parametric tests across Young Lives samples and other national samples.

The summary statistics reveal that Young Lives households were significantly poorer and had less access to basic facilities compared with the national average statistics. While household heads in the Young Lives sample were more likely to be male and more educated than in the national sample, the primary caregiver of a child is more likely to be female and less educated.

The non-parametric tests indicate that Young Lives data may not be representative at the national level but are likely to be representative at the regional level and within similar groups such as households with a 1- or an 8-year-old child.

The results reflect the sampling approach chosen by Young Lives. Due to oversampling of poor sites, it is unlikely that households from the Young Lives sample display the same characteristics than the national average. Moreover, the comparison of the Young Lives sample with nationally representative surveys is limited due to Young Lives' focus on a specific population group of households with a 1 or 8-year-old child.

The primary goal of Young Lives is to get in-depth understanding of child hood poverty and its consequences and not the provision of summative information at national level. The sample frame is geared towards these objectives. The Vietnam Young Lives sample shows that the data fits in with the project's objectives and are more representative at site level.

References

- Cameron A.C. and P.K. Trivedi (2005) *Micro econometrics: Methods and Applications*, New York: Cambridge University Press
- Committee for Population, Family and Children Vietnam (CPFCV) and ORC Macro (2003) *Vietnam Demographic and Health Survey 2002*, Calverton, MA
- General Statistics Office (GSO) (2003) *Vietnam Household Living Standard Survey 2002*, Hanoi: Statistical Publishing House
- Kanji G.K. (1993) *100 Statistical Tests*, London: Sage Publications
- National Centre for Social Science and Humanities (NCSSH) (2001) *National Human Development Report 2001: Doi moi and human development in Vietnam*, Hanoi: The Political Publishing House
- Government of Vietnam (2007) 'About Vietnam', http://www.chinhphu.vn/portal/page?_pageid=439,1090503&_dad=portal&_schema=PORTAL (accessed 15 February 2008)
- Tuan T, P. Lan, T. Harpham, N.T. Huong, D.T. Thach, B. Tod, T. Dua, N.T.V. Ha, K. Attawell (2003) *Young Lives Preliminary Country Report: Vietnam*, London: Young Lives
- Wilson I., S.R.A. Huttly, and B. Fenn (2006) A Case Study of Sample Design for Longitudinal Research: Young Lives, *Int. J. Social Research Methodology* 9(5): 351-65

Appendix 1

Young Lives study sites in Vietnam



Source: Tuan et al. 2003

Appendix 2

Non-parametric tests for equality of two distributions¹⁰

1. Kolmogorov-Smirnov two-sample test

The Kolmogorov-Smirnov test investigates the significance of the difference between two population distributions based on two sample distributions

Given samples of size n_1 and n_2 from two populations, the cumulative distribution functions $S_{n1}(x)$ and $S_{n2}(x)$ can be determined and plotted. Hence, the maximum value of the difference between the plots can be found and compared with a critical value. If the observed value exceeds the critical value, the null hypothesis that the two population distributions are identical is rejected.

2. Wilcoxon-Mann-Whitney Test

This test is used to test the null hypothesis that two random samples came from two populations with the same mean.

The results of the two samples are combined and arranged in order of increasing size and given a rank number. In cases where equal results occur, the mean of the available rank numbers is assigned. The rank sum R of the smaller sample is now found. Let N denote the size of the combined samples and n denote the size of the smaller sample. A second quantity $R' = n(N+1) - R$ is now calculated. The values R and R' are compared with critical values. If either R or R' are less than the critical value, the null hypothesis of the same mean would be rejected.

3. Median chi-square test

The Median chi-square tests the null hypothesis that two random samples could have come from two populations with the same frequency distribution.

The median of the combined sample of (n_1+n_2) elements is found. Then, for each series in turn, the number of elements above and below this median can be found and entered in a 2x2 table of the form:

	Sample 1	Sample 2	Total
Left of median	a	b	a+b
Right of median	c	d	c+d
Total	n1= (a+c)	n2= (b+d)	N=n1+n2

The test statistic is:

$$\chi^2_0 = \frac{\{|ad-bc|-1/2N\}^2 \cdot N}{(a+b)(a+c)(b+d)(c+d)}$$

If this value exceeds the critical value obtained from χ^2 tables with 1 degree of freedom, the null hypothesis of the same frequency distribution is rejected.

¹⁰ Source: Kanji 1993.

Appendix 3

Calculation of household wealth index

The wealth index is intended to be the primary instrument to measure the socioeconomic status of a household. It draws on work undertaken by the World Bank and Macro International that developed a wealth index cited in the UNICEF Multiple Indicator Cluster Surveys.

The Young Lives wealth index is constructed from three different indices: housing quality (HQ), consumer durables (CD), and services (SV). It is a simple average of these three components ($WI = (HQ+CD+SV)/3$). The calculations produce a value between 0 and 1 whereby a higher wealth index indicates a higher socioeconomic status.

Housing quality index

- HQ1: Rooms per Person. Number of rooms divided by the number of household members. The HQ1 variable is set to take a maximum value of unity. Ratios higher than 1 are recoded accordingly.
- HQ2: Floor quality. It takes the value of 1 if the floor is made of a finished material (cement, tile or laminated material); 0 otherwise.
- HQ3: Roof quality. It takes the value of 1 if the roof is made of iron, concrete tiles or slates; 0 otherwise.
- HQ4: Wall quality. Has the value 1 if the wall is made of brick or plaster and 0 if made of other materials

Consumer durables index

- Constructed from simple means of the following dummy variables:¹¹ ownership of (i) radio, (ii) bicycle, (iii) TV, (iv) motorbike or scooter, (v) motorised vehicle or truck, (vi) landline telephone, (vii) cell phone, (viii) electric fan, (ix) fridge.

Services index:

- S1: Electricity. Has value of 1 if the household has access to electricity; 0 otherwise.
- S2: Water. Has value of 1 if the household's source of drinking water is piped into dwelling or yard; 0 otherwise.
- S3: Toilet. Has value of 1 if the household has access to its own pit latrine or flush toilet; 0 otherwise.
- S4: Cooking fuel. Has value of 1 if the household uses electricity, gas or kerosene as cooking fuel; 0 otherwise.

¹¹ A dummy variable is a variable that has the value zero or one for each observation, e.g. ownership of a radio yes or no.

Appendix 4

Comparison of Young Lives 1-year-old sample with the Vietnam Household Living Standard Survey 2002

	Young Lives 2002 Sample	Vietnam Holding Living Standard Survey 2002			
	With 1-Year Old Child	National Sample		Regional Sample	
		Full Sample	With 1-Year Old Child	Full Sample	With 1-Year Old Child
Household Characteristics					
1 Average household size	4.9	4.69 ***	5.76 ***	4.49 ***	5.37 ***
2 House ownership					
Yes	76.55	96.13 ***	95.51 ***	96.65 ***	93.79 ***
No	23.45	3.87	4.49	3.35	6.21
3 Electricity supply					
Yes	84.5	84.73	79.16 ***	85.41	78.62
No	15.5	15.27	20.84	14.59	21.38
4 Source of drinking water					
Piped into dwelling/yard/plot	9.90	11.43	9.33	12.27	11.03
Tube well into dwelling/yard/plot	22.75	60.82 ***	60.59 ***	57.07 ***	52.41 ***
Public standard pipe/tubewell	0.30	4.56 ***	4.84 **	3.75	3.45
Unprotected well/spring/pond/ river/stream	40.60	17.79 ***	15.60 ***	22.82 ***	22.07 ***
Other	26.45	5.41 ***	9.63 ***	4.09 ***	11.03 ***
5 Toilet facility					
Flush toilet/septic tank	22.00	18.51 **	15.51 ***	18.28	14.48
Pit latrine (household)	26.80	24.41	19.85 ***	25.18	26.21
Pit latrine (communal)	0.35	12.89 ***	13.09 ***	21.44 ***	23.45 ***
None	35.60	18.09 ***	25.58 ***	19.12 ***	23.45 **
Other	15.25	26.10 ***	25.95 ***	15.97	12.41
6 Own radio					
Yes	43.35	28.70 ***	28.40 ***	26.32 ***	26.21 ***
No	56.65	71.30	71.60	73.68	73.79
7 Own TV					
Yes	55.35	66.44 ***	58.37	69.84 ***	60.00
No	44.65	33.56	41.63	30.16	40.00
8 Household category					
Hungry and poor	33.30	11.71 ***	13.53 ***	8.38 ***	11.72 ***
Average, better off or rich	66.70	88.29	86.47	91.62	88.28
9 Certified as poor household by local authorities					
Yes	7.00	3.70 **	4.44	2.51 **	4.14
No	26.30	4.95 ***	6.22 ***	3.60 ***	4.83 ***
Non-poor	66.70	91.35 ***	89.33 ***	93.89 ***	91.03 ***
10 Own livestock in the past 12 months					
Yes	65.55	60.79	63.56	59.24 ***	64.83
No	34.45	39.21	36.44	40.76	35.17
Number of observations	2000	29429	2025	2029	145

Note: Kolmogorov-Smirnov test results for difference between VLSS and YL variables at
***1% significance level, ** 5% significance level, *10% significance level

	Young Lives 2002 Sample	Vietnam Holding Living Standard Survey 2002			
	With 1-Year Old Child	National Sample Full Sample	With 1-Year Old Child	Regional Sample Full Sample	With 1-Year Old Child
Head of Household Characteristics					
1 Average age	37.93	47.55 ***	43.87 ***	47.53 ***	44.31 ***
2 Sex					
Male	84.35	76.38 ***	79.79 **	74.68 ***	82.07
Female	15.65	23.62	20.21	25.32	17.93
3 Education					
Not yet going to school or no qualification	31.65	40.49 ***	55.79 ***	41.18 ***	59.31 ***
Primary	30.45	25.98 ***	20.01 ***	25.17 ***	17.24 **
Secondary	23.00	20.09 *	14.83 ***	20.15	13.1
High school	8.15	7.44	5.08	7.78	6.21
Vocational school	2.70	4.08	3.06	3.50	2.76
College/university or higher	4.05	1.91	1.23	2.22	1.38
4 Marital status					
Permanent partner	97.45	81.91 ***	85.21 ***	80.05 ***	84.83 **
Divorced or separated	1.1	2.27	1.77	2.22	3.45
Single	1.05	2.24	1.23	2.41	0.69
Widowed	0.4	13.58 ***	11.78 ***	15.32 ***	11.03 *
Number of observations	2,000	29,530	2,029	2,030	145
Individual Characteristics					
1 Average age	29.54	28.57 ***	27.93 ***	28.70 ***	28.20 ***
2 Sex					
Male	46.97	49.50 ***	49.45 ***	49.45 **	49.23 *
Female	53.03	50.50	50.55	50.55	50.77
3 Education					
Not yet going to school or no qualification	30.15	41.26 ***	44.18 ***	40.85 ***	43.45 ***
Primary	36.85	26.88 ***	26.43 ***	25.68 ***	24.90 ***
Secondary	20.41	19.49	17.95 ***	19.77	18.48
High school	7.09	7.09	6.53	8.08	7.59
Vocational school	2.17	3.70 *	3.39	3.63	3.43
College/university or higher	3.33	1.59 **	1.53 **	1.99	2.15
Number of observations	7,294	132,384	75,732	8,761	5,401

Note: Kolmogorov-Smirnov test results for difference between VLSS and YL variables at
***1% significance level, ** 5% significance level, *10% significance level

Appendix 5

Comparison of Young Lives 8-year-old sample with the Vietnam Household Living Standard Survey 2002

	Young Lives 2002 Sample	Vietnam Holding Living Standard Survey 2002				
	With 8-Year Old Child	National Sample		Regional Sample		
		Full Sample	With 8-Year Old Child	Full Sample	With 8-Year Old Child	
Household Characteristics						
1	Average household size	4.93	4.69***	5.51***	4.49***	5.4**
2	House ownership					
	Yes	85.7	96.13***	96.03***	96.65***	96.52**
	No	14.3	3.87	3.97	3.35	3.48
3	Electricity supply					
	Yes	88.7	84.73*	79.91***	85.41	76.12***
	No	11.3	15.27	20.09	14.59	23.88
4	Source of drinking water					
	Piped into dwelling/yard/plot	10.9	11.43	7.84	12.27	7.96
	Tube well into dwelling/yard/plot	24.5	60.82***	63.28***	57.07***	61.19***
	Public standard pipe/tube well	0.5	4.56*	4.28	3.75	4.48***
	Unprotected well/spring/pond/river/stream	39.4	17.79***	16.03***	22.82***	15.42***
	Other	24.7	5.41***	8.56***	4.09***	10.95***
5	Toilet facility					
	Flush toilet/septic tank	20.6	18.51	14.45***	18.28	14.43
	Pit latrine (household's)	28.7	24.41*	22.99**	25.18	22.89
	Pit latrine (communal)	0.7	12.89***	11.24***	21.44***	12.94*
	None	34.4	18.09***	22.82***	19.12***	30.35
	Other	15.6	26.1***	28.51***	15.97	19.4
6	Own radio					
	Yes	42.6	28.7***	26.44***	26.32***	21.39***
	No	57.4	71.3	73.56	73.68	78.61
7	Own TV					
	Yes	69.3	66.44	62.93***	69.84	60.2
	No	30.7	33.56	37.07	30.16	39.8
8	Household category					
	Hungry and poor	33.3	11.71***	14.89***	8.38***	13.93***
	Average, better off or rich	66.7	88.29	85.11	91.62	86.07
9	Certified as poor household by local authorities					
	Yes	10	3.7***	5.14**	2.51***	3.48
	No	23.3	4.95***	6.12***	3.6***	6.97***
	Non-poor	66.7	91.35***	88.74***	93.89***	89.55***
10	Own livestock in the past 12 months					
	Yes	68.6	60.79***	66.49	59.24***	72.64
	No	31.4	39.21	33.51	40.76	27.36
Number of observations		1,000	29,429	3,480	2,029	201

Note: Kolmogorov-Smirnov test results for difference between VLSS and YL variables at

***1% significance level, ** 5% significance level, *10% significance level

	Young Lives 2002 Sample	Vietnam Holding Living Standard Survey 2002			
	With 8-Year Old Child	National Sample		Regional Sample	
		Full Sample	With 8-Year Old Child	Full Sample	With 8-Year Old Child
Head of Household Characteristics					
1 Average age	38.98	47.55***	42.1***	47.53***	39.75
2 Sex					
Male	84	76.38***	83.8	74.68***	85.07
Female	16	23.62	16.2	25.32	14.93
3 Education					
Not yet going to school or no qualification	26.30	40.49***	53.89***	41.18***	53.23***
Primary	30.60	25.98**	22.77***	25.17**	21.89
Secondary	26.50	20.09***	15.06***	20.15***	15.92**
High school	9.80	7.44	4.62**	7.78	4.48
Vocational school	3.10	4.08	2.61	3.50	3.48
College/university or higher	3.70	1.91	1.06	2.22	1.00
4 Marital status					
Permanent partner	95.5	81.91***	89.7***	80.05***	93.03
Divorced or separated	1.7	2.27	1.32	2.22	1.00
Single	0.7	2.24	1.03	2.41	0.50
Widowed	2.1	13.58***	7.94***	15.32***	5.47
Number of observations	1,000	29,532	3,487	2,030	201
Individual Characteristics					
1 Average age	26.72	28.57***	23.52***	28.70***	22.60***
2 Sex					
Male	47.99	49.50	48.9	49.45	48.88
Female	52.01	50.50	51.5	50.55	51.12
3 Education					
Not yet going to school or no qualification	26.90	41.26***	54.41***	40.85***	57.23***
Primary	33.22	26.88***	23.9***	25.68***	19.40***
Secondary	27.34	19.49***	14.35***	19.77***	14.84***
High school	8.10	7.09	4.33***	8.08	4.46
Vocational school	1.88	3.70	2.32	3.63	3.10
College/university or higher	2.57	1.59	0.67	1.99	0.97
Number of observations	3,930	132,384	18,161	8,761	1,031

Note: Kolmogorov-Smirnov test results for difference between VLSS and YL variables at

***1% significance level, ** 5% significance level, *10% significance level

Appendix 6

Comparison of Young Lives with the Demographic and Health Survey 2002

	Young Lives 2002 sample with 1-year-old child	Young Lives 2002 sample with 8-year-old child	DHS 2002 sample
Household Characteristics			
1 Average household size	4.9	4.9	4.4
2 Electricity supply			
Yes	84.5	88.7	88.4
No	15.5	11.3	11.6
3 Source of drinking water			
Piped into dwelling/yard/plot	9.9	10.9	19.4
Tube well into dwelling/yard/plot	22.8	24.5	51.9
Public standard pipe/tube well	0.3	0.5	3.2
Unprotected well/spring/pond/river/stream	40.6	39.4	24.4
Other	26.5	24.7	1.1
4 Toilet facility			
Flush toilet/septic tank	22.0	20.6	24.7
Pit latrine (household's)	26.8	28.7	56.2
Pit latrine (communal)	0.4	0.7	0.9
None	35.6	34.4	17.4
Other	15.3	15.6	0.9
5 Own radio			
Yes	43.4	42.6	49.8
No	56.7	57.4	50.2
6 Own TV			
Yes	55.4	69.3	70.0
No	44.7	30.7	30.0
Head of Household Characteristics			
1 Average age	37.9	39.0	48.2
2 Sex			
Male	84.4	84.0	72.3
Female	15.7	16.0	27.7
3 Education			
Not yet going to school or no qualification	31.7	26.3	8.8
Primary	30.5	30.6	30.6
Secondary	23.0	26.5	56.6
High school	8.2	9.8	
Vocational school	2.7	3.1	4.0
College/university or higher	4.1	3.7	
4 Marital status			
Permanent partner	97.5	95.5	89.0
Divorced or separated	1.1	1.7	0.8
Single	1.1	0.7	9.3
Widowed	0.4	2.1	0.9
Number of observations	2,000	1,000	7,048

	Young Lives 2002 sample with 1-year-old child	Young Lives 2002 sample with 8-year-old child	DHS 2002 sample
Primary Caregiver Background			
1 Literacy rate	79.9	76.7	83.91
Read easily	7.15	10	7.51
Read with difficulty	12.95	13.3	8.58
Cannot read			
2 Highest grade completed in school	11.00	10.70	6.4
No school	27.50	30.40	45.2
1-5 years (primary level)	46.55	42.10	31.5
6-9 years (secondary level)	14.95	16.70	16.8
10-12 years (high school level)			
3 Ethnic background	84.50	87.30	83.9
Kinh	15.50	12.70	16.1
Minority			
4 Religion	1.45	2.10	5.1
Christian	10.00	8.80	12
Buddhist	82.70	83.10	79.1
None	5.85	6.00	3.8
Other			
Pregnancy, Delivery and Breastfeeding			
1 Receiving antenatal care	82.0	N.A	86.0
Yes	17.1	N.A	13.0
No	0.9	N.A	1.0
Not known			
2 Number of antenatal visits during pregnancy	18.25	N.A	13.2
None	11.75	N.A	10.1
1	46.15	N.A	47.4
2-3 visits	23.85	N.A	29.3
4+ visits			
3 Receiving two or more injection for Tetanus during pregnancy	65.45	N.A	71.0
Yes	15.85	N.A	15.0
No	18.7	N.A	14.0
Other			
4 Place of delivery	18.9	N.A	19.12
Home	49.95	N.A	49.64
Hospital	28.2	N.A	26.42
Commune health centre	2.9	N.A	4.82
Other	0.05	N.A	0
Not known			
5 Breastfeeding	98.1	N.A	98.0
Yes	0.95	N.A	2.0
No	0.95	N.A	
Not known			
Number of observations	2,000	1,000	5,665

THE AUTHOR

Ngoc P. Nguyen is currently a DPhil student in Economics at the University of Oxford. Her areas of interest include childhood poverty, migration, remittances and poverty and applied micro-econometric techniques using household survey data.