YOUNG LIVES TECHNICAL NOTE **NO. 3**
March 2008

An Assessment of the Young Lives Sampling Approach in Peru

Javier Escobal
Eva Flores

Young Lives
An International Study of Childhood Poverty

Contents

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 examine how policies affect children's well-being and (iii) to inform the development and implementation of future policies and practices that will reduce childhood poverty.

In Peru the Young Lives team used multi-stage, cluster-stratified, random sampling to select the two cohorts of children. This methodology, unlike the one applied in the other Young Lives countries, randomised households within a site as well as sentinel site locations. To ensure the sustainability of the study, and for resurveying purposes, a number of well-defined sites were chosen. These were selected with a pro-poor bias, ensuring that randomly selected clusters of equal population excluded districts located in the top five per cent of the poverty map developed in 2000 by the *Fondo Nacional de Cooperacion para el Desarrollo* (FONCODES, the National Fund for Development and Social Compensation).

This paper assesses the sampling methodology by comparing the Young Lives sample with larger, nationally representative samples. The Peru team sought to:

- analyse how the Young Lives children and households compare with other children in Peru 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 Peru
- determine whether different levels of living standards are represented within the dataset.

We used two nationally representative comparison samples: the Living Standard Measurement Survey 2001 (ENAHO 2001) and the Demographic and Health Survey 2000 (DHS 2000). We used two different methodologies to assess the Young Lives sample. We first compared poverty rates calculated for Young Lives and the ENAHO 2001, then compared wealth index scores for the Young Lives households with those for DHS 2000 households. This provided a graphical illustration of the relative wealth of the Young Lives sample relative to the population of Peru. 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 2000, and the ENAHO 2001 samples. Finally, we investigated potential biases in the Young Lives and DHS 2000 samples by comparing these with the Census 2005. We compared variables that are common in the Census 2005, the DHS 2000, and Young Lives – area of residence, access to electricity and access to drinking water. In order to ensure comparability of the different samples we imposed constraints on the comparison samples to accommodate the fact that the Young Lives sample only includes households with at least one child aged between 6 and 18 months.

We found that the poverty rates of the Young Lives sample are similar to the urban and rural averages derived from the ENAHO 2001. Households in the Young Lives sample were found to be slightly wealthier than households in the DHS 2000 sample. A similar picture emerged when we use unweighted t-tests to compare the means for a range of living standard indicators between the Young Lives and the DHS 2000 samples. Young Lives households own more private assets and have better access to public services such as drinking water and electricity supply. Similarly, members of households in the Young Lives sample are better educated and have better access to vaccinations and prenatal care than DHS 2000 households. To establish the existence of biases in the Young Lives and the DHS 2000 samples we compared both with data from the Census 2005. It was evident that the Young Lives sample includes households with better access to electricity and drinking water than the Census 2005.

To reduce these noticeable biases and to improve the comparability of the Young Lives sample at national and regional levels we used post-stratification, a technique used in survey analysis to incorporate the population distribution of important characteristics into survey estimates. We post-stratified the Young Lives sample and the DHS 2000 sample against the Census 2005. Many of the differences, which we observed in the comparison of Young Lives to the DHS 2000 without post-stratification, were reduced. However, differences in access to health services and prenatal care persist.

The analyses show that households in the Young Lives sample are better-off than the average household in Peru, as measured by the nationally representative DHS 2000. However, most of the differences initially observed between the samples disappear when the sampling frames are taken into consideration. Nevertheless, households in Young Lives appear to be located in sites with better access to health, education and other services.

After using post-stratification to control for potential biases in the Young Lives sample and in the DHS 2000 sample, many differences between the samples are not significant. However, some differences between the Young Lives sample and the DHS 2000 sample remain. It is evident that post-stratification can help to better balance the Young Lives samples, especially in comparison to nationally representative samples.

In summary, we find that Young Lives households are very similar to the average household in Peru, although they may have better access to some services. Despite these biases, it is shown that the Young Lives sample covers the full diversity of children in Peru 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 and modelling child welfare, and its longitudinal dynamics in Peru.

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 has been following two cohorts in each study country. The younger cohort or one-year-old cohort consists of 2,000 children per study country aged between 6 and 18 months in 2002. The older cohort of eight-year-olds consists of 1,000 children per country who were aged between 7.5 and 8.5 years in 2002. The key objectives of Young Lives are: (i) to improve the understanding of causes and consequences of childhood poverty, (ii) to examine how policies affect children's well-being, (iii) to inform the development and implementation of future policies and practices that will reduce childhood poverty.

To expand the utility of the data generated by Young Lives and to enhance the policy impact of the research there is a need to understand the nature of the data gathered and, at the same time, to connect the data to national or regional statistics available in each country. A better understanding of how Young Lives data is related to national census data or to data from nationally representative surveys such as Living Standard Measurement Surveys (LSMS) or Demographic and Health Surveys (DHS) can provide guidance for interpreting research that uses Young Lives data.

The objectives of this report are to describe the first round sample and the sampling design of Young Lives in Peru, and to derive appropriate sampling weights needed to use the data. In addition, by comparing Young Lives with nationally representative surveys that were carried out at the same time that the Young Lives sample was collected, we provide an assessment of potential biases in the Young Lives data. Finally, the comparison between the national census data and Young Lives allows us to suggest potential post-stratification weights that may be used to adjust Young Lives sample estimates to a known population at national or regional level.

The report examines how appropriate it is to use Young Lives sample averages without considering the sampling design. The finding of this analysis might also be applicable to the three other Young Lives countries. The comparison of Young Lives data to other datasets allows us to: (i) make Young Lives comparable with national surveys carried out in the country at specific periods; (ii) identify and characterise potential biases in the Young Lives sample; and (iii) evaluate scope for adjusting Young Lives sample estimates to known population figures through post-stratification.

This report is structured as follows. In section 2, we review the sampling strategy used in the four countries to select the Young Lives samples. Then we specify the strategy used in Peru. In section 3, we compare simple, weight-adjusted averages of Young Lives data with data from the Peruvian LSMS (ENAHO 2001) from the *Instituto Nacional de Estadística e Informática* (INEI, National Institute for Statistics and Informatics) (INEI 2001a) and the Peruvian Demographic and Health Survey 2000 (DHS 2000) (INEI 2001b). The comparison allows us to assess biases in the Young Lives sample. In section 4, we compare data from Young Lives and the DHS 2000 with data from the Peruvian Household and Population Census 2005 (INEI 2006). We use a post-stratification strategy in the comparisons that may be used to adjust Young Lives sampling averages. Finally, section 5 summarises the results and defines the methodological steps needed to do similar adjustments in the other Young Lives countries.

2. Young Lives sampling strategy

Young Lives used a sentinel site sampling approach. This consisted of a multistage sampling procedure, whereby 20 sentinel sites per study country were selected non-randomly, while 100 households within a sentinel site were chosen randomly. According to Wilson et al. (2006), this strategy was thought as a way of looking at 'mini-universes' in which detailed data could be collected in order to build up a comprehensive picture of the site as well as tracking changes over time. To fit the main objectives of Young Lives, poor areas were purposively over-sampled and rich areas were excluded from the sample.

Young Lives wants to investigate characteristics of children living in poverty rather than to produce national average statistics. Therefore, the sampling approach differs from a random cluster sampling. Given this decision, the project was originally framed as:

> … 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.
>
> (Wilson et al. 2006, p. 358)

Young Lives includes 2,000 children per country who were aged between 6 and 18 months in 2002. This small sample size has a limited statistical basis. Nevertheless, the project needs to develop a way to relate to the entire population of the country from which the samples were drawn.

2.1 Sampling strategy used in Peru

While the Peru research team followed the general principles laid out by Young Lives there were some differences that affected the way the sample relates to the entire population. In Peru, the sentinel sites were chosen using a multi-stage, cluster-stratified, random sampling approach, while in the three other countries a non-random sampling approach was used.

The procedure followed in Peru was as follows: (Escobal et al. 2003)

1. The initial sample frame used in Peru was the district level. The most recent poverty map[1] of the 1,818 districts in Peru at that time was developed by *Fondo Nacional de Compensación y Desarrollo Social* (FONCODES, the National Fund for Development and Social Compensation) in 2000 and was used as the basis for selecting the 20 sentinel sites. FONCODES ranked all districts in Peru by a poverty index, which was calculated from variables such as infant mortality rates, housing, schooling, roads, and access to services. To achieve the aim of over-sampling poor areas, the five per cent of districts ranked highest were excluded from the sample. This then enabled a systematic selection of the remaining districts yielding approximately 75 per cent of sample sites considered as poor and 25 per cent as non-poor. Districts were listed in rank order with their population sizes and divided into equal population groups. A random starting point was selected and a systematic sample of districts was chosen using the population list.

[1] Poverty maps are geographical profiles that show the spatial distribution of poverty within a country, and where policies could have the greatest impact on poverty reduction.

Ten selection runs were made by computer and the resulting samples of districts were examined for their coverage of rural, urban, peri-urban, and Amazonian areas and for logistical feasibility. We chose the sample of districts that best satisfied the requirements of the study. Figure 2 in Appendix 1 shows the positions of the selected districts along the FONCODES poverty index scale.

2. Once the districts were chosen, maps were obtained from INEI of census tracts – a small geographical area that can be covered by one census worker in a short time. Census tracts can vary in size according to density of the population, geographical dispersion, and other characteristics. Using random number tables, one census tract in each district was randomly selected. All *manzanas* (blocks of houses) and *centros poblados* (clusters of houses) in the chosen census tract were counted. Using random number tables, one *manzana* or *centro poblado* was randomly selected per district.
3. All households in the selected manzanas or centros poblados were visited by a fieldworker to identify households with at least one child who was aged between 6 and 18 months in 2002. Then, the neighbouring *manzanas* or *centros poblados* were visited until a total of 100 eligible households were found. This method introduced a spatial correlation problem into the sample.

Probability of selecting a district

Since all districts were divided into equal population groups before selecting 20 sentinel sites, the probability of selecting a district was proportional to its population size. Similarly, all census tracts within a district and all manzanas or centros poblados within a census tract had the same probability of being selected. Finally, the probability of selecting a child is proportional to the average number of eligible children per household in this district. Hence, the expansion factors for the Young Lives sample in Peru can be calculated as follows:

$$f_{\exp} = \frac{Pop_T}{Pop_{d_i}} * \frac{NHou_{e_i}}{NChild_{d_i}}$$

Where Pop_T is the total population in the country; Pop_{d_i} is the total population in the selected district; $NHou_{e_i}$ the number of eligible households in the district; and, $NChild_{d_i}$ is the number of eligible children in the district.

Table 1 summarises the selected districts, the regions in which they are located, the FONCODES poverty index, the poverty ranking, the population size, the number of eligible households, the number of eligible children selected, and the expansion factor. Table 7 in Appendix 1 presents some additional information for the selected districts.

The population size of the selected districts varies strongly. Hence, the proportion of eligible children per selected district differs. We account for these differences by different weighting factor.

Table 1. *Characteristics of districts sampled in Young Lives in Peru*

District	Region	Poverty Index	Poverty classification	Population	Number of eligible households	Number of children selected	Expansion factor
1	Tumbes	15.07	average	90,625	5,350	100	522
2	Piura	21.08	poor	22,279	1,462	100	580
3	Piura	38.43	very poor	11,564	523	101	392
4	Amazonas	32.99	very poor	7,697	478	101	538
5	San Martín	30.25	very poor	16,194	1,237	101	662
6	San Martín	16.28	average	66,997	3,045	102	386
7	Cajamarca	22.35	poor	141,588	7,950	107	434
8	La Libertad	20.35	poor	124,766	7,070	102	482
9	Ancash	26.05	poor	9,585	476	103	414
10	Ancash	17.97	average	55,732	2,306	105	332
11	Huánuco	42.69	extremely poor	10,773	757	101	609
12	Lima	14.60	average	713,018	39,943	100	495
13	Lima	17.81	average	380,480	21,245	102	475
14	Lima	14.24	average	324,107	18,205	103	468
15	Junín	27.41	poor	24,376	1,839	105	605
16	Ayacucho	35.50	very poor	7,392	1,064	108	1091
17	Ayacucho	23.00	poor	17,068	1,052	102	524
18	Apurimac	28.99	poor	15,282	1,099	105	577
19	Arequipa	19.12	average	10,329	310	102	255
20	Puno	23.12	poor	189,275	10,150	102	456
Total				**2,239 127**	**125,561**	**2,052**	

Sources: FONCODES 2000 and INEI 2006

3. Potential biases in the Young Lives sample and suggested adjustments

To assess how the Young Lives sample relates to nationally representative samples we calculated poverty rates for both the Young Lives sample and the ENAHO 2001 sample. To assure comparability we narrowed the ENAHO 2001 sample down to include only households with at least one child aged below the age of one. We further excluded households from districts located in the top five per cent of the FONCODES poverty map. The ENAHO 2001 used a three-stage, stratified, random, cluster sampling approach. Since we had information about the sampling design, we adjusted the standard errors and confidence intervals of our estimations in view of the fact that we used a sub-sample of the original sample.

Table 2 compares the poverty rates calculated for the Young Lives sample and the ENAHO 2001 sample. As can be seen, the confidence intervals overlap which means that the poverty rates of the Young Lives sample are similar to the urban and rural averages derived from the nationally representative ENAHO 2001.[2]

Table 2. *Comparison of poverty rates of Young Lives and ENAHO 2001 (in %)*

	Young Lives poverty rates[a]		
	Estimate	**99% confidence interval**	
Rural (based on income)	86.3	81.0	91.5
Urban (based on income)	68.3	63.0	73.6
Total Peru (based on income)	77.3	70.3	84.3
	ENAHO 2001-IV poverty rates[b]		
	Estimate	**99% confidence interval**	
Rural (based on income)	90.0	80.0	99.9
Rural (based on expenditure)	89.6	86.4	92.8
Urban(based on expenditure)	55.9	51.1	60.7
Total (based on expenditure)	69.9	66.4	73.5

Notes: [a] For this exercise we assumed equally-proportion clustered sampling
[b] For households with at least one year-old child

Sources: Young Lives first round data and INEI 2001a

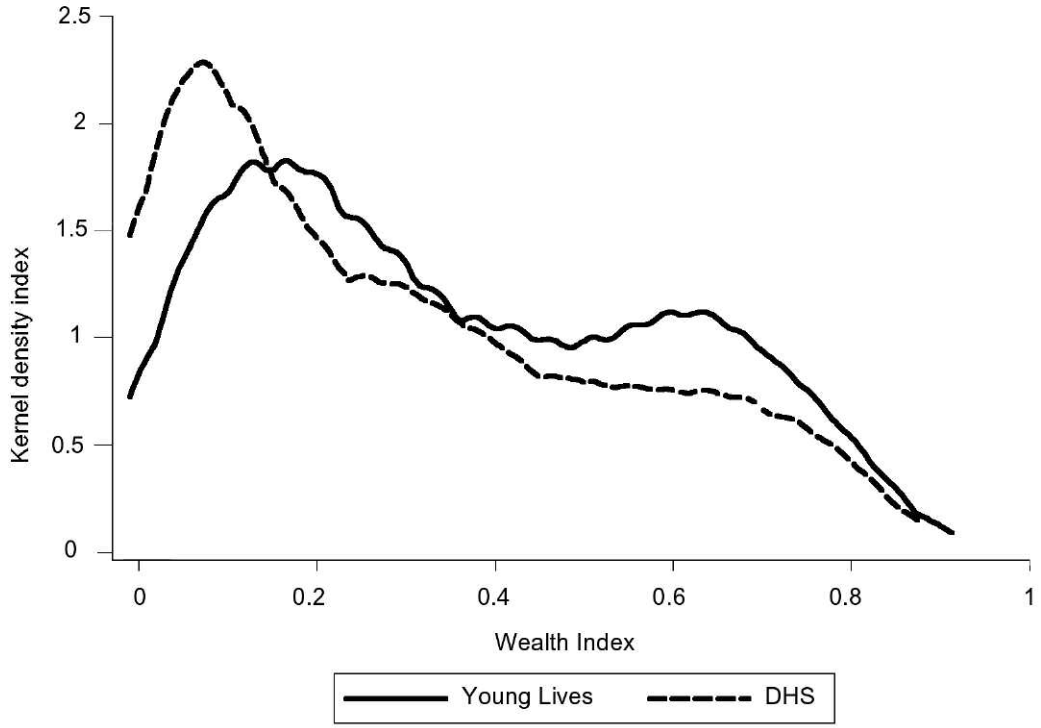
[2] A confidence interval is an interval estimate of a population parameter. For example, a 95 per cent confidence interval means that there is a 95 per cent confidence that the true population value of a variable falls within this interval.

Poverty rates cannot be calculated for all four Young Lives countries because income data were not always collected. However, in all four countries wealth index scores were calculated for each sentinel site as a measure of economic well-being. Arbitrary thresholds of the wealth index of 0.2 and 0.4 were introduced to classify the sites in the poorest, the moderately poor, and the least poor sites. This approach was justified by work undertaken by the World Bank and Macro International that developed a wealth index cited in the UNICEF Multiple Indicator Cluster Surveys (UNICEF 2007).[3] A wealth index is commonly used by countries when DHS samples are described (Filmer and Pritchett 1999). The index is designed to include sufficient variables that can vary substantially across a sample according to wealth (Filmer and Pritchett 1998).

The wealth index does not capture changes in wealth (except dramatic changes) and is therefore not a good indicator for longitudinal studies. Nevertheless, wealth indexes were calculated in round one of Young Lives. Because of this shortcoming, we did not use the arbitrary thresholds but divided the Young Lives sample into three groups based on the wealth index scores: poorest (T1), moderately poor (T2), and least poor (T3).

Figure 1 shows the wealth index distribution across the Young Lives sample and the DHS 2000 sample. The distributions are very similar; however, the Young Lives sample is slightly wealthier than the DHS 2000 sample.

Figure 1. *Wealth index distribution functions, Young Lives and DHS 2000*



[3] Macro International is a company that focuses on research and evaluation, management consulting, information technology, and social marketing communications. It provides research-based solutions, for the private and public sector and contributes to Demographic and Health Surveys.

To assure comparability between the Young Lives sample and nationally or regionally representative survey, both samples have to refer to the same population.

Demographic and Health Surveys are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition (Measure DHS 2007). They are a helpful tool for Ministries of Health and others institutions, providing reliable data on maternal and child health patterns. In Peru, only women of reproductive age (between the ages of 15 and 49 and children under the age of five years were targeted by the DHS 2000. Probability sampling was used and the sample was self-weighted by administrative department and area. It was a stratified, multi-stage, and independent sample within each region. Further information about the DHS 2000 sampling frame can be found in Appendix 2.

To assure comparability we limited the DHS 2000 sample to only include households with at least one child between 6 and 18 months and excluded households from the top five per cent of the FONCODES poverty map. See Table 3 for details of the DHS 2000 sub-sample.

Table 3. *Sub-sample of the DHS 2000*

Region	Population	Number of eligible households	Number of selected children	Correction factor
Amazonas	389,700	24,359	116	332258
Ancash	788,542	51,107	83	1085852
Apurimac	418,882	24,019	120	345080
Arequipa	601,581	50,692	45	1070930
Ayacucho	592,193	33,880	154	358036
Cajamarca	1,359,023	73,278	99	1579339
Cusco	1,116,979	62,753	101	1224672
Huancavelica	425,472	26,273	182	419986
Huánuco	705,647	40,510	129	712182
Ica	400,527	32,949	50	779227
Junín	1,090,429	52,894	93	1214519
La Libertad	1,154,962	80,206	77	1325097
Lambayeque	1,080,794	54,682	74	1270310
Lima 1[a]	3,975,923	393,296	91	2808774
Lima 2[a]	–	–	25	913382
Loreto	778,073	61,552	146	767424
Madre de Dios	83,894	5,657	115	62710
Moquegua	33,949	6,959	8	166967
Pasco	266,764	13,646	92	239195
Piura	1,526,284	91,730	90	1625586
Puno	1,237,413	57,241	125	937963
San Martín	669,973	39,723	85	705976
Tacna	105,816	12,320	22	345686
Tumbes	191,713	11,312	96	221088
Ucayali	402,445	25,884	109	371044
Total	**19,396,978**	**1,326,922**	**2,327**	

Source: INEI 2001b

Note: [a] Lima is divided into 35 districts called Lima 1 to 35.

The comparison is carried out at national, rural, urban, and regional levels. We divided the samples into three wealth segments based on wealth index scores: poorest (T1), moderately poor (T2), and least poor (T3).

Appendix 3 presents the results of the comparison of Young Lives to the DHS 2000 at national level. The table shows simple averages for key variables. Standard t-tests[4] were used to test for the statistical significance of differences between the samples. The sample frames were not taken into consideration in this comparison.

The differences between both samples are highly significant. Households in the Young Lives sample seem to have better access to private assets and public services such as electricity supply, drinking water, and sewerage. The unweighted sample averages also show that households in the Young Lives sample are slightly better educated, have greater access to health services, vaccinations, prenatal visits, and midwife services. However, children in the Young Lives sample are more likely to be underweight than children in the DHS 2000 sample.

When we include the samples frame in the calculation, differences found between the Young Lives and the DHS 2000 samples are not significant anymore. Hence, we conclude that the differences between the samples can be almost fully accounted for by the different sample frames used (see Appendix 4).

Nevertheless, some differences between the samples remain. For example, although, there is no significant difference between ownership of private assets in the samples, households in the Young Lives sample are more likely to be in areas with better access to public services. Furthermore, households in the Young Lives sample receive more prenatal and child health care than households in the DHS 2000 sample.

There are two possible explanations for these potential biases. First, we do not know whether the DHS 2000 itself has some biases. It might be that the DHS 2000 is slightly biased towards households with low access to public services and health facilities. Second, both samples could be biased with respect to census estimates. We assess potential biases of the DHS 2000 and the Young Lives sample in section 4.

[4] The standard t-test is used to establish the significance of the difference between the means of two samples.

4. Using the Census 2005 for post-stratification

To establish whether biases exist, and if so, their nature, we compare the DHS 2000 and the Young Lives sample with the most reliable information available: the Census 2005.

For comparison purposes, we limit the Census 2005 sample to only include households with at least one child aged between 6 and 18 months[5] and exclude households that are located in the top five per cent of districts in the FONCODES poverty map. We compare variables that are common in the Census 2005, the DHS 2000, and Young Lives: area of residence, access to electricity and access to drinking water. Table 4 presents the results for the Census 2005. As can be seen, 38 per cent of households in the Census 2005 sample are in rural areas, while 62 per cent are in urban areas. 43.9 per cent of households in the Census 2005 sample do not have access to electricity and 55.5 per cent do not have drinking water supply. In rural areas, 78.7 per cent of households do not have access to drinking water and 79.8 per cent have no electricity supply.

Table 4. *Area of residence and access to public services, Census 2005*

Urban (62% of households)

		Access to drinking water (%)		
Access to electricity (%)		yes	no	total
	yes	53.9	24.2	78.1
	no	4.8	17.2	21.9
	total	58.7	41.3	100.0

Rural (38% of households)

		Access to drinking water (%)		
Access to electricity (%)		yes	no	total
	yes	8.4	11.8	20.2
	no	12.9	66.9	79.8
	total	21.3	78.7	100.0

Overall access to drinking water: 44.5 % of households
Overall access to electricity: 56.1 % of households
Source: INEI 2006

[5] In the Census 2005, the age of children is not given in months but in years. Therefore, we included all households with at least one child below the age of two in the sub-sample.

Table 5 and 6 show the same variables for the Young Lives sample and the DHS 2000 sample. Comparing the three samples, it becomes evident that the Young Lives sample includes households with better access to electricity and drinking water. For example, while in the Census 2005 sample 56 per cent and in the DHS 2000 sample 49 per cent of households have access to electricity, in the Young Lives sample 60 per cent of households have electricity access. Furthermore, 44 per cent of households in the Census 2005 sample and 48 per cent of households in the DHS 2000 sample have drinking water supply, while 75 per cent of households in the Young Lives sample have access to this service. These differences could be due to biases in the Young Lives sample. Logistical feasibility and budget constraints of Young Lives meant that some better-endowed areas with better access to public services were selected.

Table 5. *Area of residence and access to public services, DHS 2000*

Urban (40% of households)

		Access to drinking water (%)		
Access to electricity (%)		**yes**	**no**	**total**
	yes	62.8	20.9	83.7
	no	4.0	12.3	16.3
	total	66.8	33.2	100.0

Rural (60% of households)

		Access to drinking water (%)		
Access to electricity (%)		**yes**	**no**	**total**
	yes	13.1	7.4	20.5
	no	19.1	60.4	79.5
	total	32.2	67.8	100.0

Overall access to drinking water: 47.9 % of households
Overall access to electricity: 49.3 % of households
Source: INEI 2001b

Table 6. *Area of residence and access to public services, Young Lives 2002*

Urban (66% of households)

		Access to drinking water (%)		
Access to electricity (%)		**yes**	**no**	**total**
	yes	75.6	9.3	85.0
	no	6.8	8.2	15.0
	total	82.4	17.6	100.0

Rural (34% of households)

		Access to drinking water (%)		
Access to electricity (%)		**yes**	**no**	**total**
	yes	16.8	5.1	21.9
	no	46.7	31.4	78.1
	total	63.5	36.5	100.0

Overall access to drinking water: 74.8 % of households
Overall access to electricity: 59.6 % of households
Source: Escobal et al. 2003

These biases in the Young Lives sample could affect our analysis, especially if we want to engage in national policy discussions.

To reduce biases in the Young Lives sample, we use post-stratification, a technique used in survey analysis to incorporate the population distribution of important characteristics into survey estimates. Post-Stratification can improve the accuracy of survey estimates both by reducing biases and by increasing precision (Zhang 2000). It may also correct for non-response bias. However, post-stratification has limitations that need to be assessed carefully. For example, if the study population is a sample drawn from the entire population, post-stratification cannot claim that the reweighed sample can approximate the entire population.

Post-Stratification takes advantage of the random clustered (and eventually, stratified) nature of the sample and combines it with complementary knowledge of the population.

In post-stratification, the sample is divided into strata based on characteristics of the population. Then individuals in each cell (post-stratum) are weighted up to the population total count for that cell. This procedure is called raking. The weights can be calculated with the following formula:

$$w_k = u_k \frac{N_{ij}}{\hat{N}_{ij}}$$

Where u_k is the sample unit weight to be modified; w_k is the modified weight; N_{ij} is the known population count for cell ij, and $\hat{N}_{ij}$ is the estimated population count for cell ij.

If modified weights are used for analysis, there is the implicit assumption of equal probability of inclusion in the sample within cells. The probability includes both design and non-response issues (Gelman and Carlin 2001).

The raking process is an iterative process that reaches a convergence according to the set of weights presented as result.[6] See Appendix 7 for more information on raking.

Some common variables that could be used for post-stratification of our samples are:

- area of residence (urban /rural)
- access to public services (electricity, piped water, sewerage)
- characteristics of the dwelling (number of rooms, type of floor, wall, roof)
- maternal characteristics (age, level of education).

We use three strata to post-stratify the Young Lives and the DHS 2000 sample against the totals obtained from the Census 2005. These strata are used to reweight the sample against area of residence, access to electricity and access to drinking water. As shown, there are some biases related to these variables. We explored if post-stratification can help to reduce these biases and further biases that might have been generated by them. Appendix 6 compares the Young Lives and the DHS 2000 sample after raking. Many of the differences, which we observed in the comparison of the samples without raking, are reduced and no longer significant. For instance, there were no significant differences in the access to sewerage using raking. The Young Lives wealth index that was higher and significantly different from the DHS 2000 wealth index before raking, was now slightly lower and non-significant. In some cases, differences between the samples disappeared entirely after raking.

Nevertheless, some differences persisted even after raking. For example, there were still significant differences in child health variables such as recently having had a fever and in the prevalence of stunting and underweight.

This exercise showed that raking or some other post-stratification technique can be used for Young Lives data and the results (with and without post-stratification) can be compared with the nationally representative sample, the sampling frames can be taken into consideration. Moreover, we could test whether the Young Lives sample averages fall within the 95 per cent confidence intervals of the nationally representative sample. However, this is not strictly needed for the Young Lives sample in Peru since we were able to construct appropriate sampling weights.

[6] If one has detailed information on the marginal cells, a variation of the raking command can be used in STATA SE v 9.1: instead of the raking option, a post option can converge the results not only at aggregate level, but also at each of the sub-group levels specified as stratum.

5. Conclusion

We have described the process we developed to compare the Young Lives sample with nationally or regionally representative surveys. We identified that a better understanding of possible, intentional, or unintentional biases is the first step for their correction. We defined the following steps to identify and address biases in the Young Lives sample:

1. Identification of a nationally representative survey (for example a DHS or a LSMS) as a comparator.
2. Sub-samples comparable with the Young Lives sample are created from the nationally representative survey and common variables with compatible definitions and categories are identified. In this case, we chose to calculate a wealth index to assess the existence of potential biases and as a comparative tool between surveys.
3. We made an initial comparison between the Young Lives sample and a nationally representative survey (in our case the DHS 2000). We used standard t-tests to test for significance of differences between the samples. We did not consider the different sample frames in our analysis.
4. We did the same comparison between the Young Lives sample and the DHS 2000 sample but considered the sample frames. Many of the significant differences observed without considering the sample frames were reduced or eliminated showing that the initial analysis might be misleading.
5. We used a post-stratification technique called raking to control for potential biases in the Young Lives sample and in the DHS 2000 sample. We used three strata to post-stratify the samples against the totals obtained from the Census 2005. These strata were used to reweight the sample. We repeated the comparison between the samples after raking. Many differences were reduced and not significant anymore. However, some differences between the Young Lives sample and the DHS 2000 sample remained after raking.

In Peru, we found that many potential differences between the Young Lives sample and a nationally representative sample could be accounted for by incorporating the sampling frames of the surveys. Raking allowed us to better balance the sample by access to private assets and public services. However, the procedure had limited power for balancing several child and mother health outcomes that continued to be different between both surveys. Finally, we see a need for further assessment of the use of post-stratification in longitudinal studies.

References

Escobal J., C. Lanata, S. Madrid, M. Penny, J. Saavedra, P. Suarez, H. Verastegui, E. Villar, S. Huttly (2003) *Young Lives Preliminary Country Report: Peru*, London: Young Lives

Gelman A. and J.B. Carlin (2001) 'Post-stratification and Weighting Adjustments' in R. Groves, D. Dillman, J. Eltinge, and R. Little (ed.) *Survey Non-response*, New York: Wiley

Filmer D. and L. Pritchett (1999) The Effect of Household Wealth on Educational Attainment: Evidence from 35 Countries, *Population and Development Review*, 25.1: 85-120

Filmer D. and L. Pritchett (1998) Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrolments in States of India, *Demography*, 38.1: 115-32

Fondo Nacional de Compensación y Desarrollo Social (FONCODES) (2001) Mapa de la Pobreza 2000, Lima: FONCODES

Instituto Nacional de Estadística e Informática (INEI) (2001a) *Encuesta Nacional de Hogares (ENAHO) 2001 IV Trimestre*, Lima

Instituto Nacional de Estadística e Informática (INEI) and ORC Macro (2001b) *Encuesta Demográfica de Salud Familiar 2000*, Lima and Calverton: INEI and ORC Macro http://www.measuredhs.com/pubs/pub_details.cfm?ID=334&ctry_id=33&SrchTp=ctry (accessed 28 April 2008)

Instituto Nacional de Estadística e Informática (INEI) (2006) *Censos Nacionales 2005*, Lima: INEI http://www.inei.gob.pe

Measure DHS (2007) *'DHS Surveys'*, http://www.measuredhs.com/aboutsurveys/dhs/start.cfm

UNICEF (2007) 'Multiple Indicator Cluster Survey-Assessing the economic status of households' http://www.childinfo.org/MICS2/finques/gj00106a.htm (accessed 28 April 2008)

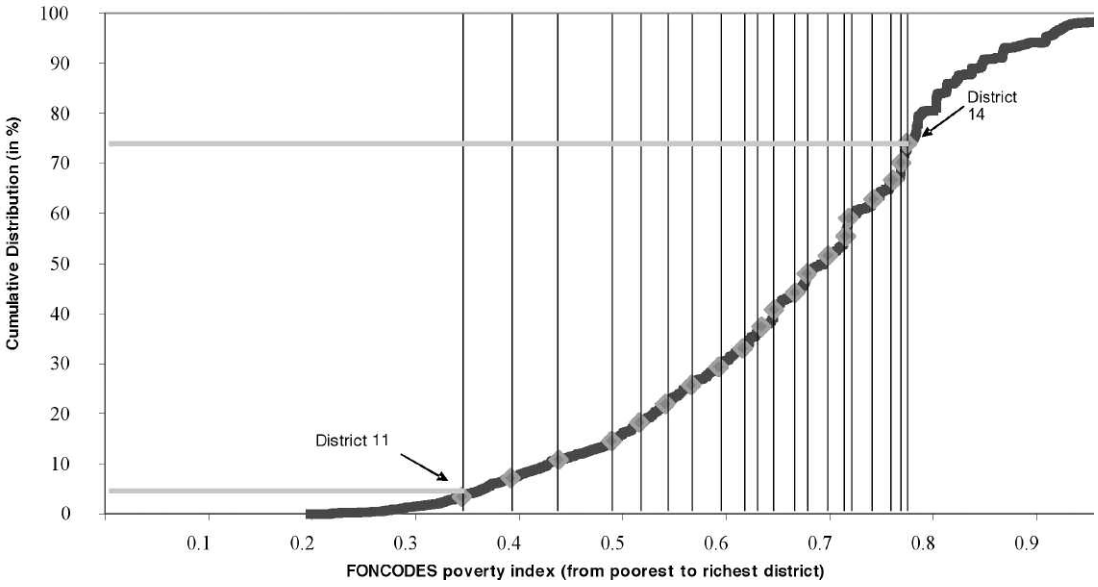
Wilson I., S.R.A Huttly, and B. Fenn (2006) 'A Case Study of Sample Design for Longitudinal Research: Young Lives', *Int. J. Social Research Methodology*, 9.3: 351-65

Zhang L.C. (2000) Post-Stratification and Calibration-A Synthesis, *The American Statistical*, 54.3:178-84

Appendix 1

Positions and characteristics of districts where Young Lives sentinel sites are located

Figure 2. *Position of the 20 selected districts along the FONCODES poverty scale*



Source: Escobal et al. 2003

Table 7. *Characteristics of selected districts in Young Lives*

District	Region	Absolute Poverty Index	Poverty classification	Poverty Rank	Population	Malnutrition rate	Without piped water access (%)	Without sewage access (%)	Without electricity access (%)
11	Huánuco	0.655	extremely poor	161	10,773	48.5	55.4	94.3	97.2
3	Piura	0.608	very poor	305	11,564	46.9	68.7	82.9	89.9
16	Ayacucho	0.562	very poor	462	7,392	46.7	0.0	99.9	95.5
4	Amazonas	0.511	very poor	662	7,697	42.8	8.3	78.2	64.9
5	San Martín	0.484	very poor	786	16,194	38.8	94.6	58.7	52.8
18	Apurimac	0.459	poor	919	15,282	47.9	62.1	81.3	37.8
15	Junín	0.434	poor	1036	24,376	44.4	32.2	89.0	63.6
9	Ancash	0.407	poor	1150	9,585	47.1	0.0	96.3	31.8
17	Ayacucho	0.384	poor	1259	17,068	38.1	31.3	99.2	32.4
20	Puno	0.366	poor	1351	189,275	24.9	31.3	99.2	32.4
7	Cajamarca	0.354	poor	1401	141,588	34.7	27.4	45.1	38.5
2	Piura	0.334	poor	1466	22,279	28.6	13.1	82.7	63.2
8	La Libertad	0.322	poor	1511	124,766	21.8	34.7	41.2	24.8
19	Arequipa	0.302	average	1564	10,329	23.6	34.3	89.7	43.4
10	Ancash	0.285	average	1623	55,732	34.5	8.8	36.9	28.0
13	Lima	0.282	average	1631	380,480	18.5	40.8	52.6	27.6
6	San Martín	0.258	average	1674	66,997	18.4	15.2	0.0	12.0
1	Tumbes	0.239	average	1702	90,625	16.9	22.2	43.2	22.6
12	Lima	0.231	average	1712	713,018	17.4	43.1	48.5	23.5
14	Lima	0.282	average	1726	324,107	14.8	34.1	39.0	22.9
Total					**2,239,127**				

Appendix 2

Sampling procedure for the Demographic and Health Survey 2000

Probability sampling was used in the DHS 2000. The sample was selected in three stages.

1. Populated centres (cities, town, villages, etc.) were systematically selected with probability proportional to size sampling. They represent the primary sampling units (PSU).
2. The PSU were divided into clusters of houses and clusters were selected. They were the secondary sampling units (SSU).
3. SSU were divided into dwellings and some were selected as the tertiary sampling units (TSU). The selection assured the same sample fraction for dwellings in each department.

The sample consisted of 1,414 clusters, allocated proportionally within urban and rural areas in each department of Peru (see Table 8). The INEI keeps a list of all populated centres in Peru. They are stratified into urban centre, suburban area, and rural areas. There was an average of 50 clusters per department, with the exception of Lima, where 226 clusters were sampled. The sample is self-weighted by department. For estimates at national level appropriate correction factors must be applied for each department.

Source: INEI 2001b, Appendix A

Table 8. *Selected clusters in the DHS 2000*

Department	Number of women interviewed	Urban centre	Suburban area	Rural area	Total
Amazonas	1,000	5	9	36	50
Ancash	1,100	13	14	25	52
Apurimac	1,000	7	6	37	50
Arequipa	1,150	40	5	11	56
Ayacucho	1,200	14	8	38	60
Cajamarca	900	4	6	40	50
Cusco	900	10	7	33	50
Huancavelica	1,000	6	2	42	50
Huánuco	1,000	13	3	34	50
Ica	1,000	28	12	10	50
Junín	1,100	17	13	20	50
La Libertad	1,200	23	12	17	52
Lambayeque	1,000	26	12	12	50
Lima	3,600	198	13	15	226
Loreto	1,200	27	8	23	58
Madre de Dios	1,000	23	2	25	50
Moquegua	1,000	31	5	14	50
Pasco	1,000	14	9	27	50
Piura	900	9	21	20	50
Puno	1,200	15	8	37	60
San Martín	1,000	12	15	23	50
Tacna	1,000	40	2	8	50
Tumbes	1,000	24	17	9	50
Ucayali	1,000	28	6	16	50
Total	**27,450**	**627**	**215**	**572**	**1,414**

Source: INEI 2001b

Appendix 3

Comparison of Young Lives to the DHS 2000 at national level without sample frame

(using wealth index groups (T1-T3), at national level, in %)

Socioeconomic status of the household

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Main material of the floor: cement, ceramic tiles or vinyl	0.0	0.5	6.8	23.7	68.9	88.5	25.1	36.9	***
Main wall material: bare bricks or cement blocks	0.0	0.4	2.7	11.5	52.8	83.8	18.4	31.4	***
Main roof material: concrete or tiles	0.0	9.5	34.0	22.3	45.0	66.3	25.9	32.4	***

Access to public services

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Source of drinking water: piped water into dwelling	5.0	50.1	43.9	85.9	85.1	97.3	44.2	77.0	***
Type of toilet facility: flush toilet at home	0.1	4.0	3.6	39.2	54.3	92.0	19.3	44.1	***
Access to electricity	2.5	20.1	39.0	80.3	94.4	99.4	44.8	65.2	***

Household assets

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Own fridge	0.0	0.5	3.3	8.0	37.5	42.4	13.5	16.7	***
Own radio	52.1	59.5	82.0	80.5	94.4	84.3	75.8	74.3	
Own TV	7.0	15.7	42.6	70.9	91.2	92.3	46.5	58.4	***
Own car	0.5	0.0	1.2	2.3	12.3	10.0	4.6	4.0	
Own phone	0.0	0.0	0.1	1.4	19.1	23.3	6.4	8.1	**
Type of cooking fuel: gas or electricity	0.2	1.2	8.7	25.5	52.7	76.8	20.4	33.8	***
Wealth index	0.0385	0.1167	0.1994	0.3478	0.5635	0.7121	0.2649	0.3860	***

Respondent characteristics

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Average age	26.9	26.8	27.4	27.2	27.9	27.5	27.4	27.2	
Level of Education									
None	14.4	15.7	11.2	6.7	1.8	0.6	9.2	7.8	
Primary school.	62.6	58.1	53.3	38.7	22.9	12.9	46.4	37.1	***
Secondary school	20.0	23.5	30.3	42.9	46.6	49.9	32.2	38.3	***
Higher education	3.0	2.3	5.2	10.5	28.7	36.3	12.2	16.1	***
Obese (BMI > 30)[a]	6.1	4.9	8.1	7.7	13.8	16.8	9.3	9.8	
Overweight (BMI 25.0-29.9)[a]	31.1	27.2	37.0	40.7	50.9	54.1	39.6	40.3	
Marital status									
Single	5.8	10.0	7.5	7.7	8.8	9.5	7.3	9.1	**
Married	31.3	39.6	39.8	35.1	37.7	31.0	36.2	35.3	
Living together	31.3	46.4	39.8	50.8	37.7	53.9	36.2	50.3	
Current pregnant	3.2	2.2	2.1	1.5	2.3	1.9	2.6	1.9	

Pregnancy and delivery

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Pregnancy not wanted	54.0	48.2	52.7	46.9	47.3	43.7	51.4	46.3	***
Received prenatal care	48.8	89.8	55.0	90.9	72.7	95.6	58.7	92.1	***
Received tetanus injection during pregnancy	43.9	71.8	47.0	72.3	60.6	75.2	50.5	73.1	***
Assistance during delivery									
Doctor	11.3	30.1	19.6	49.1	46.6	66.5	25.7	48.1	***
Nurse	10.3	21.2	13.2	25.5	17.3	22.0	13.6	22.8	***
Other birth attendant	6.8	3.0	8.9	3.6	16.7	5.5	10.8	4.0	***
Partner	32.7	23.1	27.8	9.6	9.6	2.4	23.4	12.0	***
Relative	33.5	19.9	26.2	10.1	8.3	2.8	22.7	11.2	***
Place of delivery									
Home	74.0	55.8	59.3	26.9	20.8	7.7	51.6	30.8	***
Hospital	12.7	21.0	22.7	46.4	58.3	72.4	31.1	46.0	***
Other health facility	12.4	21.2	16.2	24.1	19.5	18.3	16.0	21.2	***
Other	0.9	1.9	1.1	2.5	0.5	1.3	0.8	1.9	***
Had caesarean section	3.4	5.8	6.4	11.5	16.9	19.8	8.9	12.2	***
Ever had an abortion	10.2	15.7	9.3	20.3	14.1	29.1	11.2	21.6	***

Child health

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Sex – male	51.1	47.3	52.1	51.7	49.0	51.4	50.7	50.0	
Average birth weight	3128.1	3119.9	3119.2	3196.1	3230.5	3274.4	3159.3	3194.9	*
Ever breastfed	98.6	99.3	98.3	98.6	96.9	97.2	97.9	98.4	
Ever vaccinated	91.9	97.4	94.3	98.1	96.3	98.7	94.1	98.0	***
Ever had health card	90.2	97.7	92.9	97.8	96.8	97.0	93.3	97.5	***
Had fever in the previous two weeks	30.7	12.9	30.2	10.7	25.7	7.4	28.9	10.4	***
Had cough in the previous two weeks	43.2	44.6	42.1	41.6	44.6	39.0	43.3	41.8	
Had rapid breathing in the previous two weeks	21.9	7.3	20.2	5.3	17.7	3.4	19.9	5.4	***
Stunting	30.5	37.0	27.4	23.5	10.6	10.8	22.9	24.1	
Underweight	12.4	17.0	10.0	9.9	4.0	3.7	8.9	10.4	*

Source: Young Lives and INEI 2001b

Notes: [a] Body mass index is an individual's weight divided by the square of their height

*Full sample differences are significant at 10%; **significant at 5%; ***significant at 1%

Appendix 4

Comparison of Young Lives and the DHS 2000 with sample frame

(using DHS sample frame, at national level, in %)

Socioeconomic status of the household

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Main material of the floor: cement, ceramic tiles or vinyl	32.4	30.1	27.1	33.1	
Main material of the wall: bare bricks or cement blocks	27.4	23.9	21.1	26.9	**
Main material of the roof: concrete or tiles	29.5	30.0	26.9	33.2	

Access to public services

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Source of drinking water: piped water into dwelling	74.8	47.9	44.3	51.5	***
Type of toilet facility: flush toilet at home	38.3	23.7	20.6	26.7	***
Access to electricity	59.6	49.2	45.6	52.9	***

Household assets

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Own fridge	14.5	15.5	13.2	17.7	
Own radio	74.3	78.0	75.7	80.3	***
Own TV	53.8	52.0	48.7	55.3	
Own car	3.5	5.8	4.5	7.2	***
Own phone	6.9	8.9	7.1	10.8	**
Type of cooking fuel: gas or electricity	29.5	24.8	22.1	27.5	***

Respondent characteristics

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Average age	27.1	27.5	27.1	27.9	**
Level of Education					
None	9.3	8.8	7.25	10.4	
Primary school	40.8	43.3	40.43	46.3	*
Secondary school	35.5	35.3	32.5	38.1	
Higher education	13.7	12.6	10.7	14.5	
Obese (BMI > 30)[a]	8.9	10.2	8.6	11.9	
Overweight (BMI 25.0-29.9)[a]	37.9	41.3	38.6	43.9	**
Marital status					
Single	9.4	7.5	6.2	8.9	***
Married	36.5	36.2	33.5	38.9	
Living together	48.9	50.1	47.4	52.8	
Current pregnant	1.8	2.9	1.9	3.7	***

Pregnancy and delivery

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Pregnancy not wanted	46.2	52.0	49.4	54.6	***
Received prenatal care	91.9	60.1	57.4	62.7	***
Received tetanus injection during pregnancy	74.4	51.6	48.9	54.3	***
Assistance during delivery					
Doctor	45.1	28.5	25.9	31.1	***
Nurse	23.2	12.1	10.3	13.9	***
Other birth attendant	3.9	11.1	9.3	12.9	***
Partner	12.4	24.0	21.4	26.7	***
Relative	13.3	21.0	18.6	23.5	***
Place of delivery					
Home	34.4	49.9	46.7	52.9	***
Hospital	41.2	31.8	28.9	34.6	***
Other health facility	22.3	16.4	14.3	18.5	***
other	1.9	1.1	0.5	1.6365	***
Had caesarean section	11.2	10.1	8.5	11.8	
Ever had an abortion	19.8	12.0	10.3	13.7	***

Child health

Variables	Full sample		95% Confidence interval DHS		
	Young Lives	DHS	Lower bound	Upper bound	
Sex – male	49.4	51.6	49.1	54.2	*
Average birth weight	3170.2	3187.2	31.7	3244.0	
Ever breastfed	98.6	98.0	97.4	98.7	
Ever vaccinated	98.1	94.4	93.2	95.5	***
Ever had health card	97.6	93.6	92.2	94.9	***
Had fever in the previous two weeks	11.1	28.2	25.9	30.4	***
Had cough in the previous two weeks	42.7	43.4	41.0	45.8	
Had rapid breathing in the previous two weeks	5.5	19.9	18.0	21.9	***
Stunting	27.0	15.2	13.4	17.1	***
Underweight	11.2	8.2	6.9	9.5	***

Source: Young Lives and INEI 2001b

Notes: [a] Body mass index is an individual's weight divided by the square of their height

*Full sample differences are significant at 10%; **significant at 5%; ***significant at 1%

Appendix 5

Comparison of Young Lives and the DHS 2000 with sample frame and wealth index groups

(using sample frame, wealth index groups (T1-T3), at national level, in %)

Socioeconomic status of the household

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Main material of the floor: cement, ceramic tiles or vinyl	0.0	0.3	12.5	14.2	79.1	83.6	30.1	32.4
Main material of the wall: bare bricks or cement blocks	0.6	0.0	5.4	7.4	66.4	75.7	23.9	27.5
Main material of the roof: concrete or tiles	4.1	6.1	40.4	25.0	49.0	58.3	30.0	29.5

Access to public services

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Source of drinking water: piped water into dwelling	13.7	45.6	46.2	84.1	87.0	96.9	47.9	74.8 ***
Type of toilet facility: flush toilet at home	0.4	2.5	4.8	25.2	66.5	88.3	23.7	38.3
Access to electricity	6.0	11.5	50.3	72.5	95.7	98.1	49.2	59.6 ***

Household assets

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Own fridge	0.0	0.4	5.0	5.0	42.0	38.3	15.5	14.5
Own radio	55.3	62.0	88.5	77.7	93.1	84.2	78.0	74.3
Own TV	8.3	14.3	58.0	57.7	94.3	91.9	52.0	53.8
Own car	0.4	0.0	2.5	1.9	14.9	8.6	5.8	3.5
Own phone	0.0	0.0	0.5	0.6	26.4	20.1	8.9	6.9
Type of cooking fuel: gas or electricity	1.2	0.4	11.9	16.3	62.6	72.6	24.8	29.5
Wealth index	0.0568	0.1010	0.2451	0.2970	0.6753	0.6753	0.3021	0.3541

Respondent characteristics

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Average age	27.6	26.6	27.1	27.3	27.9	27.3	27.5	27.1
Level of education								
None	16.8	17.4	6.6	9.7	1.9	0.5	8.8	9.4
Primary school	61.9	61.3	50.1	45.7	16.8	14.6	43.3	40.8
Secondary school	17.8	18.9	36.8	36.4	53.1	52.1	35.3	35.5
Higher	3.5	2.0	6.5	7.1	28.1	32.3	12.6	13.7
Obese (BMI > 30)[a]	6.6	4.6	8.6	6.9	15.8	15.1	10.2	8.8
Overweight (BMI 25.0-29.9)[a]	32.8	24.8	39.7	38.1	51.9	51.5	41.3	37.9
Marital status								
Single	5.5	11.4	7.2	7.3	10.2	9.4	7.5	9.5
Married	33.4	39.3	37.6	41.8	37.9	28.6	36.2	36.5
Living together	53.7	45.2	49.8	46.4	46.5	55.4	50.1	49.0
Current Pregnancy	3.2	1.6	2.7	2.1	2.7	1.8	2.9	1.8

Pregnancy and delivery

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Pregnancy not wanted	54.2	46.9	53.6	47.7	48.2	44.2	52.0	46.3	*
Received prenatal care	48.3	90.9	57.3	89.3	75.5	95.5	60.1	91.9	***
Received tetanus injection during pregnancy	43.2	73.6	47.9	73.7	64.3	75.8	51.6	74.4	***
Assistance during delivery									
Doctor	11.6	28.0	23.4	43.3	51.7	64.7	28.5	45.1	**
Nurse	7.7	20.3	12.6	26.1	16.4	23.5	12.1	23.2	***
Other birth attendant	6.5	3.2	8.9	3.5	18.0	5.2	11.1	3.9	***
Partner	36.3	22.8	27.3	11.3	7.6	2.5	24.0	12.4	**
Relatives	32.2	22.6	25.4	13.5	4.8	3.4	21.0	13.3	**
Place of delivery									
Home	75.5	59.7	57.3	33.7	14.9	8.6	49.9	34.5	*
Hospital	13.5	18.0	25.2	37.2	57.7	69.3	31.8	41.1	
Other health facility	9.4	20.2	15.6	26.7	24.9	20.3	16.4	22.3	
Other	1.1	2.0	1.1	2.4	1.0	1.5	1.1	1.9	
Had caesarean section	3.4	5.4	8.0	9.7	19.4	18.5	10.1	11.2	
Ever had an abortion	10.1	13.9	9.7	17.9	16.1	27.8	12.0	19.8	**

Child health

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Sex – male	52.8	46.4	49.1	52.6	52.7	49.3	51.6	49.3	
Average birth weight	3142.4	3062.1	3122.3	3182.2	3295.9	3273.2	3187.2	3170.3	
Ever breastfed	98.8	99.3	97.6	98.9	97.7	97.4	98.0	98.6	
Ever vaccinated	91.5	97.2	95.4	98.5	96.6	98.5	94.4	98.1	***
Ever had health card	89.6	97.8	94.7	98.2	97.0	96.8	93.6	97.6	***
Had fever in the previous two weeks	33.0	14.5	27.6	11.1	23.5	7.5	28.2	11.1	***
Had cough in the previous two weeks	45.2	45.1	42.3	42.6	42.5	40.1	43.4	42.7	
Had rapid breathing in the previous two weeks	23.1	7.7	18.2	5.5	18.0	3.4	19.9	5.5	***
Stunting	31.4	38.8	25.7	28.7	6.2	11.4	21.3	26.5	
Underweight	12.3	17.0	9.4	12.9	2.6	3.5	8.2	11.2	

Source: Young Lives and INEI 2001b

Notes: [a] Body mass index is an individual's weight divided by the square of their height

*Full sample differences are significant at 10%; **significant at 5%; ***significant at 1%

Appendix 6

Comparison of Young Lives and the DHS 2000 with raking weights

(using raking weights, wealth index groups (T1-T3), at national level, in %)

Socioeconomic status of the household

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Main material of the floor: cement, ceramic tiles or vinyl	0.0	0.7	19.5	9.9	85.9	75.0	34.7	28.1
Main material of the wall: bare bricks or cement blocks	1.0	0.0	12.1	3.5	73.9	65.4	28.7	22.7
Main material of the roof: concrete or tiles	3.8	2.9	30.8	24.2	54.9	46.4	29.5	23.9

Access to public services

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Source of drinking water: piped water into dwelling	8.5	13.4	40.6	42.9	85.6	79.5	44.5	44.5
Type of toilet facility: flush toilet at home	0.6	3.1	10.4	12.3	70.4	76.7	26.9	30.3
Access to electricity	8.4	13.4	64.6	60.8	97.1	97.8	56.1	56.1

Household assets

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Own fridge	0.1	0.0	8.8	2.8	47.3	32.2	18.5	11.5	
Own radio	52.2	55.2	88.0	81.5	93.5	82.1	77.5	72.3	
Own TV	13.0	7.1	67.5	57.9	96.0	88.2	58.2	49.7	
Own car	0.6	0.0	2.6	0.2	17.3	7.8	6.8	2.6	***
Own phone	0.0	0.0	2.1	0.1	31.7	15.3	11.1	5.1	**
Type of cooking fuel: gas or electricity	1.7	0.3	20.6	12.7	68.5	65.5	30.0	25.7	
Wealth index	0.0573	0.0650	0.2757	0.2315	0.6604	0.6012	0.3279	0.2943	

Respondent characteristics

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample	
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**
Average age	27.6	26.6	26.8	26.9	28.3	27.0	27.6	26.8
Level of education								
None	15.0	18.3	4.6	9.9	1.5	1.3	7.1	10.1
Primary school	57.7	57.7	40.0	52.7	14.8	19.1	37.8	43.4
Secondary school	22.3	21.4	47.4	32.0	52.9	52.6	40.6	35.0
Higher	4.9	2.4	8.0	4.3	30.8	26.7	14.5	11.0
Obese (BMI > 30)[a]	8.0	5.5	9.9	7.6	16.5	13.5	11.4	8.8
Overweight (BMI 25.0-29.9)[a]	34.6	25.0	41.7	34.1	53.3	50.7	43.1	36.5
Marital status								
Single	6.5	11.2	7.6	7.3	9.5	9.4	7.8	9.4
Married	30.0	37.5	34.9	38.8	39.8	29.8	34.8	35.4
Living together	54.0	48.1	50.9	49.7	45.7	54.2	50.2	50.6
Current Pregnancy	3.2	1.6	2.9	2.9	2.5	1.3	2.9	1.9

Pregnancy and delivery

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Pregnancy not wanted	55.1	46.7	53.1	47.9	46.7	44.4	51.7	46.3	
Received prenatal care	51.1	89.3	61.3	90.0	77.3	94.2	63.1	91.2	***
Received tetanus injection during pregnancy	45.9	72.1	51.9	75.3	65.2	74.7	54.2	73.9	***
Assistance during delivery									
Doctor	15.2	26.3	31.7	40.4	54.1	62.0	33.5	42.5	
Nurse	8.3	20.6	17.3	24.8	15.5	24.3	13.6	23.1	***
Other birth attendant	8.6	3.6	10.0	4.2	19.6	4.8	12.7	4.2	***
Partner	34.9	23.5	20.6	11.5	6.6	2.7	20.9	12.9	***
Relatives	28.3	22.7	18.4	15.5	2.8	5.6	16.6	14.8	
Place of delivery									
Home	68.4	61.3	42.9	35.1	12.2	11.8	41.5	36.8	
Hospital	19.2	18.9	36.8	36.0	61.0	65.5	38.8	39.6	
Other health facility	10.1	18.4	18.3	24.3	24.8	20.8	17.6	21.0	**
Other	1.5	1.5	0.7	4.0	1.2	1.8	1.1	2.4	*
Had caesarean section	4.4	5.7	11.1	10.3	20.5	17.9	11.9	11.2	
Ever had an abortion	11.3	16.1	10.2	15.6	17.2	26.4	12.9	19.4	*

Child health

	T1 (Poorest)		T2 (Moderately poor)		T3 (Least poor)		Full sample		
Variables	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	**DHS**	**Young Lives**	
Sex – male	52.7	44.9	48.5	54.0	52.0	48.8	51.1	49.0	
Average birth weight	3105.8	3094.1	3174.9	3166.2	3299.0	3278.6	3192.2	3177.6	
Ever breastfed	98.5	99.2	97.9	98.9	97.5	98.0	98.0	98.7	
Ever vaccinated	92.5	96.4	96.1	98.2	96.4	98.3	95.0	97.6	*
Ever had health card	89.9	97.1	95.8	98.7	96.7	96.1	94.1	97.3	**
Had fever in the previous two weeks	33.7	15.8	24.6	12.0	24.1	6.9	27.5	11.7	***
Had cough in the previous two weeks	47.5	47.8	42.0	49.5	43.4	41.5	44.4	46.3	
Had has rapid breathing in the previous fortnight	23.5	8.1	18.4	7.8	18.2	3.5	20.1	6.5	
Stunting	29.5	40.1	20.9	29.3	4.9	13.2	18.6	27.9	*
Underweight	12.3	15.9	7.4	15.0	2.3	3.1	7.4	11.4	*

Source: Young Lives and INEI 2001b

Notes: [a] Body mass index is an individual's weight divided by the square of their height

*Full sample differences are significant at 10%; **significant at 5%; ***significant at 1%

Appendix 7

Raking

Raking works as an iterative process whereby applying $w_k = u_k \frac{N_{ij}}{\hat{N}_{ij}}$, a convergence is reach into a new set of weights according to the distribution gave by the total marginal of a bigger survey, in our case, gave by the Census 2005.

For example, we stratify our sample (N=150) into sub-groups of gender and self-reported race:

	White	Black	Total
Women	36	46	**82**
Men	34	34	**68**
Total	**70**	**80**	**150**

The marginal totals in the sample are for women 82, men 68, white 70, and black 80. There are no missing data, thus both marginals sum up to the sample total.

To use raking we need to have the control totals of both variables for the population, in our case the Census 2005 information:

Women	2000
Men	2500
White	3000
Black	1500

The first step is to divide the universe one stratum (in this case, gender) into sample total, obtaining new multiplying factors by which the original sample can be multiplied:

	Factor		
Women	24.3902	Men	36.7647

	White	**Black**	**Total**
Women	878.05	1121.95	2000.00
Men	1250.00	1250.00	2500.00
Total	2128.05	2371.95	4500.00

As we can see, the objective of stratifying by gender has been reached, but the second strata still presents some discrepancies. Raking repeats this process, but taking the second stratum universe totals this time, meaning, the variable “race”, where the new results are applied into the new set of frequencies obtain from the first weighting exercise:

	Factor		
White	1.4097	Black	0.6324

	White	**Black**	**Total**
Women	1237.82	709.51	1947.33
Men	1762.18	790.49	2552.67
Total	3000.00	1500.00	4500.00

In this round, the second stratum now not tallies exactly to Census 2005 totals. However, the first stratum does not tally exactly to the Census 2005 as it did previously. It is for this reason that raking is an iterative process that repeats itself until a convergence is reached. The next round will produce the following table:

	Factor		
Women	1.0270	Men	0.9794

	White	**Black**	**Total**
Women	1271.30	728.70	2000.00
Men	1725.82	774.18	2500.00
Total	2997.12	1502.88	4500.00

Now we can see that the first stratum fits perfectly, but the second does not. Considering this example, convergence would be reached very rapidly. After convergence the next results are:

	Factor		
Women	1.0001	Men	0.9999

	White	**Black**	**Total**
Women	1272.63	727.37	2000.00
Men	1727.36	772.64	2500.00
Total	2999.99	1500.01	4500.00

The table shows the new distribution of frequencies obtained after raking; where a reliability of the universe information would allow us post-stratify the sample.

This procedure can be achieved through the command “survwgt” in STATA version 9.

AUTHORS

Eva Flores is a research assistant on poverty and equity at GRADE.

Javier Escobal is Principal Investigator for Young Lives in Peru and Senior Researcher at GRADE where his work focuses on poverty and rural development.

Young Lives
An International Study of Childhood Poverty

www.younglives.org.uk

Department of International Development
University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK
Tel: +44 (0)1865 289966
Email: younglives@younglives.org.uk