Young Lives

Section 5: Sampling

May 2017

The overall objective of Young Lives research is to produce detailed, long-term panel data about the causes and consequences of childhood poverty, the impact of pro-poor policies, and the means by which poverty is transmitted across generations in Ethiopia, India, Peru and Vietnam.

In a cohort study collecting data about the same group of people over a specified period of time, initial decisions about sample selection are a crucial determinant of the outcome of the research and the ways the data can be used.

Key considerations and challenges in sample design

Designing a sampling strategy for Young Lives involved striking a balance between many competing needs (Wilson, Huttly and Fenn 2006). Perhaps the most basic of these was the tension between selecting a statistically viable sample which was not only within the study's budget but also feasible to manage given the geographic and infrastructural characteristics of the four countries and the degree of detail demanded by the research objectives. As a result, the Young Lives sampling method was never designed to be nationally representative of children of a specific age, as achieving this within the available budget would have meant limiting the number of countries in the study. Instead, the sampling method was intended to generate a large enough sample for general statistical analysis, and to be systematic and clearly justified. This has shaped the character of Young Lives as 'an in-depth study of relationships between pieces of information, rather than an instrument to collect national statistical results' (Wilson, Huttly and Fenn 2006: 358).

The objectives of Young Lives, established prior to the sample design stage, were particularly important in shaping the approach taken to sampling. Studying the causes and consequences of childhood poverty meant designing a sample that included a high proportion of poor children, but which also included other children with whom their experiences could be compared. This was achieved by over-sampling poor areas, and then randomly selecting children of the right age within the selected communities. Avoiding a sample comprised exclusively of poor children not only provided opportunities to compare poor and better-off children, but also minimised the chance that the study results would be rejected on the grounds of not being representative.

The sample also had to be suitable for use to obtain data about children's experiences of poverty at different levels, including the community and the household. This need for detailed site-level data, together with the logistical considerations that arise from widely dispersed rural populations poorly served by transport infrastructure, determined that children would be selected in geographically compact sites rather than randomly across countries. As well as being predominantly located in poor areas, these sites were selected to reflect heterogeneity of ethnicity and religion in country populations.

These two needs – to over-sample the poor and to produce in-depth data about sites as well as children – were reconciled through the development of a multi-stage sampling procedure, adapted from sentinel site monitoring methods. The concept of sentinel site monitoring comes from public health studies, and involves the purposive sampling of a small number of settings, deemed to represent a certain type of population or area, which are then studied in a consistent way at relatively long intervals. Under the sentinel site monitoring system adopted by Young Lives:

- sentinel sites in each study country were selected non-randomly, with rich areas excluded from the sample and poor areas purposively over-sampled.
- children in the right age group in the selected sites were sampled randomly.

Implemented in 2002, this procedure resulted in the random selection of 2,000 infants (aged between 6 and 18 months) living in 20 sites mostly located in poor areas of each country. At the same time, 1,000 older children (aged 7 to 8 years) were also randomly selected in the same sites.[1] Initially, work with these older children was intended to be limited to the testing of instruments and methods for later use with the younger children. Subsequently, however, the decision was taken to retain the older cohort because of the value of inter-cohort analysis which provides unique information about changes over time. As such, the two age cohorts of children form the panel for the Young Lives longitudinal survey rounds, as well as the foundation from which sub-samples for other elements of Young Lives – such as the qualitative research and school surveys – were later drawn.

Sentinel site selection

For each country, site selection protocols were written to transparently describe the sequence of decisions that were made in selecting and defining sites and to systematise procedures for over-sampling poor areas. Proposed criteria and procedures for site selection were extensively discussed

[1] Because of resource constraints, only 700 older children were selected in Peru.

with the national Young Lives Advisory Panels and amended according to these discussions. Each of the country study teams used slightly different processes to arrive at a non-random selection of sites. Each process involved several stages.

In Ethiopia (see Outes-Leon and Sanchez 2008):

- Five regions were selected out of a total of nine, accounting for 96 per cent of the national population.
- Three to five districts were selected in each region, with a balanced representation of food-deficient rural and urban districts. Where official statistics were not available, this classification was made through consultation with local officials.
- Since districts were too large, in terms of both area and population, to be considered as sentinel sites, at least one peasant association or *kebele* (the lowest level of administration in rural and urban areas respectively) per district was selected as a sentinel site, with the key criterion being the possibility of finding at least 100 households with a 1-year-old child and 50 households with an 8-year-old child.
- A village was randomly selected within each sentinel site.

In Andhra Pradesh in India (see Kumra 2008):

- Site selection aimed to ensure a uniform distribution of sample districts across the state's three agro-climatic regions, and the inclusion of at least one poor and one non-poor district from each region.
- In order to make this selection, districts were classified and ranked according to a relative development index which aggregated economic, human development and infrastructure indicators. A representative group of 12 poor and non-poor districts was chosen from a total of 23, covering 28 per cent of the population of the state.
- *Mandals*, administrative areas containing between 20 and 40 villages, were deemed to be the appropriate size to be sentinel sites. The second step of sampling was choosing mandals within the selected districts. All the mandals in each district were ranked and selected based on a second set of economic, human development and infrastructural indicators constructed using available mandal-level data.
- Each mandal was divided into four contiguous geographical areas and one village was randomly selected from each.

In Vietnam (see Nguyen 2008):

- Five out of a total of nine provinces were selected to over-emphasise poor regions and to ensure even coverage of urban, rural and mountainous areas, and of the north, central and southern regions. The selection was made through a process of iterative consultation with a range of different actors including government, donors and NGOs.
- Working groups of provincial government staff were established to select sentinel sites in each province. All communes in each province were ranked by poverty level according the degree of infrastructural development, the percentage of poor households, and child malnutrition status. As well as level of poverty, other criteria included commitment to the research from local government officials, logistical feasibility, and adequate population to constitute a sample of children of the right age.
- Four communes were selected as sentinel sites in each selected province, 48 per cent from those ranked as poor, 29 per cent from those ranked as average and 23 per cent from those ranked as above average.

In Peru, while the research team followed the general principles of sampling agreed for the whole study, there were significant differences in sample design. Here, the sentinel sites were chosen using a multi-stage, cluster-stratified, random sampling approach (see Escobal and Flores 2008).

- Sentinel sites in Peru are districts, of which there were 1,818 at the time of sampling. A national poverty map developed in 2000 by the *Fondo Nacional de Compensación y Desarrollo Social* (National Fund for Compensation and Social Development) was used as the basis for site selection. This map ranked all districts according to a poverty index calculated from variables which included infant mortality rates, housing, schooling, roads and access to services.
- To achieve over-sampling of poor areas, the 5 per cent of highest-ranking districts were excluded from the sampling process. The remaining 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. 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. The sample of districts that best satisfied the requirements of the study was selected.
- Maps of census tracts (small geographical areas that can be covered by one census worker in a short time) were obtained for each of the selected districts, and one tract per district was selected using random number tables. In each selected tract, all *manzanas* (blocks of housing) and *centros poblados* (clusters of housing) were counted, and one was randomly selected for each district.

Child selection

Having selected 20 sentinel sites in poor areas, households containing children in the right age groups were randomly selected. While the exact procedures used by each study team were adapted to local circumstances, there was careful and transparent documentation of protocols to ensure:

- cost-effective field procedures for traversing each site.
- reasonable control of biases, for example due to the unavailability of any respondent from a household during the listing sweep through the site.
- a sample equivalent to one drawn at random from all possible qualifying households in the area (Wilson, Huttly and Fenn 2006).

In some cases, the local procedure required an exhaustive screening sweep through an administrative area like a sub-district to create a numbered list of all qualifying households, and then drawing a random sample from this list. In other cases, where a defined area was to be sampled rather than fully covered, the process included a stage adapted to the geography of households. In some densely populated urban areas, for example, this entailed selecting particular streets or alleyways as sub-units for seeking qualifying households. In some sparsely populated areas, by contrast, it entailed the use of line transects, which involved walking in a straight line between identifiable landmarks and selecting all households within 50 metres of the line (Wilson, Huttly and Fenn 2006).

The approach in each of the four countries was as follows:

- In Ethiopia, a village within each sentinel site was randomly selected and all the households on the periphery were interviewed until 150 eligible households were located.
- In Andhra Pradesh, a door-to-door listing schedule was completed in order to identify eligible children.
- In Vietnam, a door-to-door screening survey for children the right age was carried out in each commune, and simple random sampling applied to the list.
- In Peru, all households in each selected manzana or centro poblado were visited by fieldworkers to identify children of the right age. If not enough children were found using this method, then neighbouring manzanas and centros poblados were visited until the total was achieved

The Young Lives sample and national datasets

Although the Young Lives sample is not and was never intended to be nationally representative, it is important to understand how it compares with larger samples from other studies and surveys which are. In 2008, each Young Lives country sample was compared with one or two other samples to examine and discuss differences and highlight both expected and unexpected biases. This was an important step in situating the Young Lives samples in broader national contexts, and understanding what inferences could be drawn from the findings of the study.

- The Ethiopian sample was compared with the 2000 Demographic and Health Survey (DHS) and the 2000 Welfare Monitoring Survey. The analyses showed that households in the Young Lives sample were slightly better-off and had better access to basic services than the average household in Ethiopia, but that they held less land, owned less livestock, and were less likely to own a house (Outes-Leon and Sanchez 2008).
- The Andhra Pradesh sample was compared with the 1998/9 DHS. The analysis showed that households in the Young Lives sample were slightly wealthier than households in the DHS sample. They had better access to public services and owned more assets, but they were less likely to own their own house, and the mothers of Young Lives children were less likely to breastfeed or to have received an antenatal visit (Kumra 2008).
- The Vietnam sample was compared with the 2002 DHS and the 2002 Vietnam Household Living Standard Survey. The analysis showed that households in the Young Lives sample were slightly poorer than the households in the other samples. They owned fewer assets, were less likely to own their own house, and were more likely to be registered as poor by their local authorities (Nguyen 2008).
- The Peru sample was compared with the 2000 DHS, the 2001 Peru Living Standard Measurement Survey (LSMS) and the 2005 National Census. The analysis showed that the poverty rates of the Young Lives sample were similar to the urban and rural averages derived from the LSMS, and slightly wealthier than households in the DHS. Young Lives households owned more assets and had better access to public services such as electricity and drinking water than households in the other surveys (Escobal and Flores 2008).

In all four cases, analysis showed that despite biases, the Young Lives sample covered the diversity of children in each country. Therefore, while not suited for simple monitoring of child outcome indicators, the Young Lives study is an appropriate and valuable instrument for analysing causal relations, and modelling child welfare and its longitudinal dynamics.

REFERENCES

Galab, S., P.P. Reddy, and R. Singh (2014) Young Lives Survey Design and Sampling in Andhra Pradesh, India, Oxford: Young Lives.

Pankhurst,A., and T. Woldehanna (2014) Young Lives Survey Design and Sampling in Ethiopia, Oxford: Young Lives.

Sanchez, A., M. Penny, and M. Lizama (2015) Young Lives Survey Design and Sampling in Peru, Oxford: Young Lives.

Thang, N. and L.T. Duc (2014) Young Lives Survey Design and Sampling in Vietnam, Oxford: Young Lives.

Wilson, I., S. Huttly and B. Fenn (2006) 'A Case Study of Sample Design for Longitudinal Research: Young Lives', *International Journal of Social Research Methodology* 9.3: 351-365.

Young Lives

www.younglives.org.uk

Young Lives is a 15-year study of childhood poverty in Ethiopia, India, Peru and Vietnam, core-funded by UK aid from the Department for International Development (DFID).

Young Lives, Oxford Department of International Development (ODID)
3 Mansfield Road, Oxford OX1 3TB, UK
Tel: +44 (0)1865 281751 • Email: younglives@younglives.org.uk