STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 433—464

SURVEYING CHILD LABOUR THROUGH HOUSEHOLDS: SAMPLING ISSUES AND STRATEGIES

Vijay Verma[1], Francesca Gagliardi[2]

ABSTRACT

This paper addresses sampling issues arising in the context of household-based child labour surveys. It presents some of the sampling strategies elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008). A typology of surveys of child labour is identified, and the fundamental distinction between two types with very different objectives – termed 'child labour surveys' and 'labouring children surveys', respectively – is clarified and emphasised. Following a broad survey of national practices in conducting surveys of child labour, linkages between different types of surveys and some specific sampling techniques are explained.

Key words: child labour, household surveys, sampling, ILO.

1. Introduction

Child labour is an important global issue. Detailed and up-to-date statistics on working children are needed to determine the magnitude and nature of the problem, identify the factors behind child labour, reveal its consequences, generate public awareness of the related constellation of issues, and to formulate policies and projects to combat it (International Labour Organisation, 2004).

Data on child labour may be obtained from diverse sources, often used in combination. Although few *national population censuses* provide data on the prevalence of child labour, information from censuses serves as an essential basis for the interpretation and analysis of data on child labour from other sources. The population census is also the basic source of sampling frames for child labour and similar surveys. Countries collect socio-economic and demographic data through *general household-based sample surveys*, such as surveys on the labour force, living conditions, household income and expenditure, demography and health. Such surveys normally do not produce detailed data on child labour, but they can yield information that is useful for analysis of the situation concerning child

[1] University of Siena, verma@unisi.it.
[2] University of Siena, gagliardi10@unisi.it.

labour. Moreover, attaching child labour modules to such household-based surveys is also a potential source of information. In addition, a wide range of *secondary and administrative sources*, while not primarily concerned with child labour, can provide useful information pertaining to it.

Nevertheless, more comprehensive and pertinent information on child labour requires special studies and surveys focussed on the subject. Apart from *household-based child labour surveys*, with which we are concerned here, different types of instruments include rapid assessments, establishment surveys, school-based surveys, community-level enquiries, street children surveys, and baseline studies undertaken in the context of specific intervention projects. The various sources are generally complementary. The practical implication of this from the point of survey and sampling design is that the household-based instrument need not be developed to meet all the information needs: in fact some of these needs are better (or even, can only be) met from other types of instruments. Experience demonstrates that collecting comprehensive data on child labour is a challenging task, and no single survey method may in itself satisfy all data needs. "Children are found working in a vast array of circumstances, and no single technique can be devised to survey all of these situations. Furthermore, policy analysis and targeted project intervention require information from a variety of potential respondents who may influence the life and development path of the child. These include the children themselves, parents or guardians, employers, school teachers, community leaders, child peers, and siblings. Circumstances in the home, school, workplace, and larger community to which the child belongs all bear on child labour outcomes and characteristics. To collect all relevant data from all relevant parties by means of a single survey or on a single occasion is impossible." (ILO, 2004). This applies with particular force in surveying what are called the *worst forms of child labour*. It is practically impossible to make contact with children engaged in such forms of labour to collect the necessary information. Worst forms of child labour usually remain hidden and the necessary sampling frames do not exist for their enumeration. Nor can the required samples be designed and selected without prior information on the location, characteristics and circumstances of the children engaged in it. Consequently, regular household-based surveys are largely ineffective for this purpose; special sampling and enumeration procedures must be employed.

This paper is concerned with sampling issues arising in the context of household-based child labour surveys. It presents, albeit briefly and selectively, the sampling strategies elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008). This book has been reviewed in *Statistics in Transition* by Kordos (2008).

As noted, for some purposes and in certain circumstances child labour surveys may also involve non-household based data collection, and may even force some departures from the principles of probability sampling. This paper does not aim to address special considerations involved in the design of surveys aimed at estimating the prevalence and nature of child labour confined to particular sectors

or activities. Various statistical techniques used for sampling non-standard units need separate treatment, though many of the techniques discussed below can be useful in the design of such surveys as well.

2. Household-based surveys of child labour: a typology

2.1. Household-based surveys

Regular child labour surveys are household-based national sample surveys whose target are children, and also their parents or guardians living in the same household. Such surveys may be conducted as stand-alone surveys, or as separate but linked operations, or simply as modules attached to other national household-based surveys such as a labour force survey (LFS). The statistics generated by these surveys include economic activities and non-economic activities (such as household chores) of children, working hours, nature of the tasks performed, health and safety issues including injuries at work, and also background variables such as demographic and social characteristics of household members and other basic characteristics of the household.

Household-based sampling provide an efficient approach for estimating the prevalence and characteristics of predominant forms of child labour for children living in private households, irrespective of whether the work is performed at home or outside. In so far the survey is based on a scientifically designed probability sample, it permits generalization of the study results to the whole population that was sampled.

In practice, the household survey content may be detailed and specialised, providing information on the dynamics of child labour or gross flows between different child labour categories; or it may be confined to a few basic characteristics of working children. The choice depends on the data needs, available resources, and the arrangements and circumstances under which the survey is conducted. The key respondents in a child labour survey are the working and potentially working children and their parents or guardians.

Household surveys may apply a variety of designs and organizational structures. The main factors determining the design are, of course, substantive objectives concerning the content, complexity and periodicity of the information sought. These substantive requirements determine features of the survey structure such as its timing, frequency, reference period and sampling arrangements. For instance, the survey may be a continuing survey designed to obtain regular time-series data or, as has been more often the case in national surveys, it may be an occasional survey primarily for obtaining benchmark and structural information.

In the designs followed most commonly for the labour force and similar population-based surveys, especially in developing countries, the sample is selected in two (or more) stages: the selection of area units; followed by the selection of addresses, households or individuals in each area as the ultimate

sampling units. For convenience and concreteness of the exposition, we will assume such a 'typical' design throughout in the following technical discussion of sampling issues.

2.2. Child labour survey (CLS)

As a generic term we use 'a (regular household-based) child labour survey' to indicate a household sample survey the main objective of which is to provide information on the phenomenon of child labour – its prevalence, distribution, forms, economic sectors etc., as well as its conditions, characteristics and consequences. The prefix 'regular' is used to emphasise that the context is that of a broad household-based survey, as distinct from other types of studies concerning children not residing in – or at least not identified for enumeration through – private households.

While retaining the above more general, descriptive use of the term 'child labour surveys', it is very useful to keep in mind two quite different types of such surveys. These differ in their objectives, or at least in the emphasis given to different types of objectives.

The first type refers to surveys where the primary objective is to measure the *prevalence* of child labour. The surveys may also study variations in this prevalence by geographical location, type of place (urban-rural), household type and characteristics, the household's employment and income situation, children's age and gender, and similar factors. The target population of a survey with this type objectives is *the total population of children exposed to the risk of child labour*. This base population is defined essentially in terms of age limits, and therefore tends to be well-distributed in the general population. The size and structure of the sample is determined largely by the size and distribution of the population of all children, or more commonly by its approximation – the size and distribution of the general population.

We propose to reserve a more strict use of the term *Child Labour Survey* (CLS) for such surveys with the primary objective of measuring the prevalence of child labour, as distinguished from a *Labouring Children Survey* (LCS) described below. The defining factor in this distinction is the relevant base population for which the survey estimates are generated – essentially, all children within specified age limits for the CLS, and only those considered to be in child labour for the LCS.

2.3. Labouring children survey (LCS)

We have a different type of survey when the primary objective is to investigate *circumstances, characteristics and consequences of child labour*: what types of children are engaged in work-related activities, what types of work children do, the circumstances and conditions under which children work, the

effect of work on their education, health, physical and moral development, and so on. The objectives may also include investigating the immediate causes and consequences of children falling into labour. We refer to this type of survey as a *Labouring Children Survey* (LCS). The relevant base population in the LCS is *the population of working children*. What is meant by the LCS concept is that, when the objective is to determine the conditions and consequences of child labour, as distinct from its prevalence among all children, then it is appropriate that the size and structure of the sample is determined primarily by the size and distribution of the population of working children.

At the same time, it is important to clarify that the concept of a 'labouring children survey' does not imply that the ultimate units enumerated in the survey be only labouring children. On the contrary, it will normally be necessary in such a survey to enumerate comparable groups of children not engaged in labour, so as to provide a control group for comparison with the characteristics and circumstances of those subject to child labour. Nevertheless, the sample size and design of a LCS is determined primarily by the need to represent the population of labouring children; any sample of non-labouring children is supplementary, selected and added to the main sample as necessary for analytical purposes.

2.4. CLS vs. LCS: sampling implications

There are some important differences in terms of sampling aspects between Child Labour Surveys and Labouring Children Surveys in the above sense.

The target population of the CLS, being all children in a certain age range, tends to be distributed in a way very similar to the general population. Hence the required structure and distribution of the CLS sample is likely to be quite similar to that of a survey of the general population, in particular to that of the Labour Force Survey (LFS) to which the CLS is very similar in concepts, definitions and even survey content.

The target population of the LCS – whether the population of children engaged in any work-related activity, or defined more narrowly as these engaged in specific forms of child labour – is, by comparison, smaller and more unevenly distributed, often in areas of heavy concentration. Consequently, the sample design required is generally also different from that of LFS or CLS.

The two types of surveys differ in their size and complexity. The CLS is normally less intensive (i.e., involves a simpler and shorter survey interview), and requires larger sample sizes. The primary statistical consideration dictating its sample size is the precision with which the proportions of children engaged in child labour is to be estimated and the reporting domains requiring separate estimation.

By contrast, for investigating the detailed conditions and consequences of child labour, the LCS is more intensive in data collection, often involving interviewing the guardians as well as the children concerned separately, collecting attitudinal and other qualitative data, and carrying out associated enquiries such as

at the school or place of work of children in the sample. Consequently, the appropriate sample size for a LCS is likely to be much smaller than that for a CLS in similar circumstances. For an intensive survey such as the LCS, large sample sizes are often unnecessary from the statistical point of view, and in any case are precluded by practical and cost considerations. Having too large a sample in an intensive survey can in fact damage the quality and value of the information collected, in so far as it hinders close control over the survey operation.

2.5. Labouring children vs. child activity vs. children's surveys

In practice, the target populations of interest can be more diverse than the all children versus labouring children distinction of CLS and LCS.

The focus of all LCS's is on the details of child labour, or more generally, on children's work-related activities. However, some surveys collect a broader range of information on children. The following three types of situations may be distinguished.

Option (1). A majority of the LCS surveys have as their main focus the study of conditions and consequences of child labour (see Section 5 for a review of country practices).

Option (2). In a number of countries, the scope is broader, and covers all types of activities of children, including economic and non-economic activities, education, leisure, and even non-activity. Many such national surveys are actually named a *Child Activity Survey* (CAS).

Option (3). Occasionally, the scope is even broader to include more general information about children beyond their economic and non-economic activity – such as information on children's health, housing conditions, etc. These are termed a *Children's Survey*.

There is also much debate as to the definition of what constitutes 'child labour', and in particular whether 'substantial' domestic chores should be included (ILO, 2004, Chapter 2).

These variations have important sampling implications. For surveys focused on working children, option (1), the LCS samples should reflect closely the patterns of concentration of child labour. With their broader and more defused scope children's surveys, option (3), would require a design similar to that of the CLS, or may even incorporate the latter. However, even for this type of surveys, and especially for the child activity surveys, option (2), the measurement of conditions of child labour as such is likely to remain a special objective. A compromise design is therefore desirable – which covers both working and non-working children but with greater weight given to the former. Such a compromise design would of course require a CLS-type operation preceding it, so as to identify – even if with limited precision – the level of child labour in survey areas.

2.6. An example

The diverse objectives of a survey on child labour, and the diverse population groups to which the results from the survey apply, are well-illustrated by a survey from Portugal Ministry of Labour and Solidarity (1998). The survey identifies seven different target groups of questions. For three of these groups, namely

(1) families and the work of children,
(2) children and their activities in general, and
(3) aspects of children's life

the target population is all children (and their families). The questionnaire modules pertaining to these constitute the CLS component of the survey. Group (1) is the basis for estimating the proportion of working children. For another three groups, namely

(4) characteristics of children with economic activity,
(5) characterisation of those responsible for children with economic activity, and
(6) attitudes and perceptions towards child labour of the child, and of adults responsible for the child

the target population is labouring children, defined here as children engaged in any economic activity. The questionnaire modules pertaining to these constitute the LCS component of the survey. Group (4) – children engaged in economic activity – is the basis; the population of persons responsible for children in groups (5) and (6) is defined through association with the base population in group (4). The group

(7) children who carry out domestic chores

covers a somewhat different group of children. This group is often considered less important than group (4) in determining the LCS sample size requirements.

The final sample size is normally a compromise between the requirements of the CLS and LCS components. It is of course possible, in principle, to introduce sub-sampling from group (1) to group (4) (i.e., to follow-up only a sub-sample of the labouring children identified from the former); or to introduce sub-sampling from group (4) to groups (5)–(6) (i.e., to follow-up only a sub-sample of adults responsible for the labouring children). There is greater flexibility in this respect when the various components involved are operationally separated from each other.

3. Linkage of CLS to a base survey

Surveys may be needed to collect different types of information on child labour. As noted, different types of data differ in their content, mode of collection, and – with particular relevance to sampling – in the base population to which they relate and the required sample size. *Different types of data involved may be viewed as constituting different components of the survey.* The components may,

for instance, be combined into a single integrated whole; or they may remain distinct but linked in various ways; or they may form more or less separate stand-alone operations, each like a survey in itself. Similarly, the child labour survey or its components may be related in various ways to other existing surveys, such as a labour force survey. In designing the sample for a household-based child labour survey, the first step is therefore to define the survey structure, i.e., choose the manner in which different components of the survey are to be arranged in relation to each other.

In this section we consider the linkage of the CLS with the operation preceding it. There are three dimensions: (i) the preceding operation may be a household listing operation, or it may be a large-scale survey such as the LFS; (ii) the CLS may be combined with that operation, or be conducted subsequently as a separate operation; or (iii) the CLS may be conducted on the same sample, or on a sub-sample of the preceding operation. The combinations of these are shown below. Of course some of these combinations are more likely (more meaningful, practical) than others.

Link of a CLS with the operation preceding it

<table>
<tr><td colspan="5">Whether the two operations are integrated or separate</td></tr>
<tr><td>Preceding operation</td><td colspan="2">Integrated</td><td colspan="2">Separate</td></tr>
<tr><td rowspan="2">Household listing</td><td rowspan="2">CLS involving brief screening questions</td><td>same (full) sample</td><td rowspan="2">Stand-alone CLS</td><td>using the full listing sample (unlikely option)</td></tr>
<tr><td>sub-sampling (unlikely option)</td><td>sub-sampling</td></tr>
<tr><td rowspan="2">LFS or similar</td><td rowspan="2">Modular or ‘combined’ CLS</td><td>same (full) sample (used frequently)</td><td rowspan="2">Linked CLS</td><td>same (full) sample (used sometimes)</td></tr>
<tr><td>sub-sampling (used sometimes)</td><td>sub-sampling (used frequently)</td></tr>
</table>

3.1.Modular or ‘combined’ CLS

The collection of information on child labour in conjunction with a broad-based survey of the general population such as the labour force survey (LFS) may typically take the form of child labour questions attached to the LFS as a *module*. In this case, the essential CLS information, namely estimates of proportions of children in various categories who are engaged in child labour, may be obtained by extending downwards the lower age limit for the standard LFS questions on economic activity. This is a possibility when child labour is defined in terms of the standard LFS concept of economic activity (on the latter, see Hussmanns *et al*, 1990). Different and generally more elaborate questioning will be required with a different interpretation of what is meant by ‘child labour’.

The major attraction of a modular child labour survey is that it provides an economical and convenient arrangement for obtaining essential information on child labour. Furthermore, items enumerated in the base survey are available for use as explanatory and classification variables in the analysis of child labour data. Modular surveys present some potential problems, however. The number and detail of child labour items that may reasonably be inserted into an operation primarily concerned with other topics is quite limited. Secondly, in order to ensure high-quality data the various survey topics must be compatible in terms of concepts, definitions, survey methods, reference periods, coverage, and design requirements. This compatibility requirement may enforce compromises that limit the usefulness of the resulting data.

When a set of child labour questions is attached as a module to an existing base survey, it is generally understood that the number of additional questions involved is small enough not to significantly affect the base survey sample design or data collection. We use the term *combined survey* to refer to a more comprehensive version of the modular survey. It indicates the situation when the child labour questions constitute a substantial addition to the base survey, influencing the survey design, sample size and data collection operations of the latter.

Any of the above arrangements involves operational integration of the CLS with the base survey. This normally implies that the CLS module is applied on the same sample of household as the base survey. Sub-sampling at the area level (i.e. introducing the CLS as a module only in a subsample of LFS areas) can however be possible. A particularly convenient form of this is to include the CLS as a module during only some of the rounds of a continuing LFS.

3.2. Linked CLS and its sampling aspects

A *linked CLS* means a survey dependent on the base survey (such as a LFS) for its sample and possibly also for other information fed forward, but otherwise operationally separated from the latter for the purpose of data collection. A linked survey permits more detailed measurement of child labour than is normally possible in a modular survey. Also, more elaborate sub-sampling from the base survey, both at the area and the household levels, is possible. But of course there is the extra cost of separate operations.

As to the sampling aspects of such linkages, one extreme option is to draw the CLS sample as a *sub-sample of the individual children enumerated* in the base survey. At the other extreme, the two surveys may be based on *independent samples*. There are a number of intermediate possibilities.

Even when the two surveys are based on independent samples it is desirable and efficient to draw them from a *common frame or 'master sample' of area units*. This permits the sharing of costs of preparation and maintenance of the area frame.

A closer link between the surveys is obtained by basing them on a *common sample of areas*. In principle, all the areas in the base survey may be included for the CLS sample. However, sub-sampling of the LFS areas is often desirable and appropriate: the sample sizes required for a CLS tend to be much smaller than those for major surveys such as the LFS.

Within the common sample areas, various possibilities exist in terms of the relationship between the ultimate units (e.g. households) in the two samples – from entirely independent samples from household lists in the common areas, to confining the CLS to a sample of children actually identified during the LFS information on household composition.

(1) At the one extreme, the common LFS-CLS sample areas may be 're-listed' to obtain a more up-to-date frame of households for the CLS, and an entirely new sample of the units selected. However, creation or updating of household lists can be an expensive operation, and is justified only if the two surveys are separated by, say, one year or longer.

(2) When the same lists are used, the two samples may still be selected independently or without overlap. This is desirable when respondent fatigue is a concern, or when the first sample is subject to high rates of non-response. Independent or at least additional sampling is also required when the first survey is not able to yield a sufficiently large sample for the CLS.

An alternative is to base the CLS on all or a sub-sample of ultimate units included in the first sample. The outcome depends on the type and characteristics of the units involved.

(3) Selecting a sub-sample of *addresses* in the base sample is often the simplest.

(4) The use of *households* as units for sub-sampling comes next. It is simpler if households from the first survey are subject to sub-sampling without reference to any particular characteristics of the households involved.

(5) However, sometimes information on various *household characteristics* is evoked for the purpose of stratification or for applying different sampling rates. This involves the collection of such information in the base survey, and its preservation and transfer to the CLS. This can be expensive and cumbersome, and in many cases not very effective in improving the efficiency of the resulting sample. It is also common to exclude certain types of households from the selection, such as households not found to contain any child relevant to the CLS. This can improve control over the CLS sample size, and also efficiency of its fieldwork. However, the drawback is the assumption that the situation of the households with respect to the exclusion criteria has not changed during the interval between the two surveys.

(6) Using *children* identified during the first survey as the units for sampling for the CLS is also an option. However, this is a demanding choice in that lists of children identified have to be prepared, transferred to the CLS

operations for sampling, and then the selected children identified during fieldwork. Misidentification of individual children can easily occur. Such an option should be followed only if the interval between the two surveys is very short.

(7) At the extreme is the procedure where information on various *characteristics of children* is used for the purpose of stratification or for applying different sampling rates, such as the child's educational and/or activity status. Such a procedure may appear attractive when the CLS sample size is very small and its structure needs to be tightly controlled. However, generally this is a demanding and expensive procedure, prone to implementation errors. It should be used only when the CLS approximates the conditions of being a 'module' of the first survey – close in timing the first survey, and drawing significantly on the substantive information collected in it.

3.3. CLS based on a household listing operation

A child labour survey may be based on a household listing operation required to create or update the sampling frame. (In an area-based sample, such a listing operation is normally confined to sample areas selected at the preceding stages.) At least two options are possible. The child labour survey may be a separate *stand-alone* survey, conducted subsequently to a household listing operation, and not linked to another survey such as the LFS. Sub-sampling of households or similar units from the lists will be normally required within each sample area. However, a stand-alone CLS is not always a feasible or even a desirable option.

An alternative is *integration* of the CLS with household listing in the form of a single operation. This necessitates that the child labour component involves no more than a few brief questions which can be incorporated into the listing form. In an area-based sample, the listing operation normally involves exhaustive coverage of each sample area, i.e. the listing of all households or similar units in it without involving any sub-sampling. The same applies in the case of the CLS where it is operationally integrated with household listing. The resulting data may have the advantage of being based on a large sample, but the measurement of child labour is likely to be approximate. This type of CLS is essentially no more than a *screening operation* for a more detailed survey of labouring children.

4. CLS-LCS linkage

A labouring children survey (LCS) requires a prior operation which identifies, explicitly or implicitly, a sample of working children. Normally this is the function of the CLS component. Two forms of the relationship between CLS and LCS may be identified: (i) an integrated CLS-LCS operation; and (ii) linked CLS and LCS operations.

4.1. Integrated CLS-LCS operation

The distinction, discussed in Section 2, between the two types of surveys, the CLS and the LCS, by no means implies that they must (or sometimes even can) be organised as two separate operations. In fact, in a majority of the national surveys conducted so far they have been completely integrated into a single survey operation – the same survey covering the two different types of objectives. The LCS questions form additions to the CLS questionnaire, which become applicable if the child concerned is found to be engaged in work-related activity.

This is an *integrated design*, by which is meant that the information for the CLS and LCS components is collected as a single interview operation. Note that a 'single interview operation' is not meant to necessarily imply that only a single integrated questionnaire is used, or that all interviewing takes place during a single visit to the household, or even that all the information is obtained from a single respondent in the household. Multiple questionnaires, repeated interviewer visits, and different respondents (head of household, parents or guardians, children themselves, and sometimes even employers and teachers etc.) may indeed be involved. Rather, the term is meant to indicate that the information on the CLS and LCS components is collected at the same time or at least within a short time of each other, and that all the information collected during the CLS component is directly available to (and is generally not repeated in) the LCS component.

Simultaneous implementation of the CLS and LCS components implies that normally no sub-sampling from the one to the other can be introduced (i.e., all children identified as working during the CLS part are subject to the LCS part of the interview). In any case, it is desirable in practice that any sub-sampling involved is straightforward, such as applying the LCS part of the interview to only a pre-selected subset of CLS survey rounds or sample areas.

In an integrated survey, the sample size requirements of both the CLS and LCS components have to be met. Hence even in a single integrated survey, the distinction between the two types of survey components still remains a conceptually useful one: it reminds us that an integrated design has to be a compromise between different objectives. It comes to requiring that the CLS is large enough to provide estimates of prevalence of child labour with the required precision, and also to yield sufficient numbers of working children for the LCS.

Unfortunately in national practices, often the survey design has been determined one-sidedly – with over-emphasis on one type of objective at the expense of the other. For instance, some surveys have been too small to yield useful estimates of the extent and distribution of prevailing child labour (the CLS component was too small in sample size), while others have been too large in size to permit sufficiently in-depth investigation of the characteristics and consequences of child labour (LCS component was too large in sample size from the practical point of view). By contrast, there are also examples where the sample size, while adequate for the CLS, turned out to be inadequate in providing enough cases for the LCS component for the purpose of investigating child labour activities in detail.

It is worth emphasising that with an integrated design, the sample for the LCS component is determined entirely by the results of the screening provided by the CLS component: not only *sample size* but also – and more importantly – the *quality of coverage* of the population of working children in the LCS component is determined by the quality of the screening questions in the CLS part for identifying those children.

Advantages and disadvantages of an integrated design

There can be practical and cost advantages in integrating the CLS and LCS components into a single operation. Clearly, it is cheaper and convenient to collect all the required information in one go. This is a major advantage, and may explain why a large majority of child labour surveys to-date have chosen the integrated arrangement.

However, there are also disadvantages of an integrated (CLS+LCS) design.

(1) An integrated implementation tends to limit the information which can be collected during the LCS component without jeopardising the quality of the CLS component.
(2) The increased burden associated with the LCS part can have serious consequences for the quality of the CLS part in identifying the incidence of child labour.
(3) Apart from the obvious upper limit imposed by the number of eligible cases (working children) identified in the CLS component, there is no independent control over the size and distribution of the sample for which the LCS information is collected.
(4) The level of child labour may be very unevenly distributed over sample areas, thus greatly varying interview workload. This becomes all the more troublesome when the LCS part involves lengthy interviews with adults as well as with children individually.

When an integrated CLS-LCS operation may be suitable

An integrated CLS-LCS operation may well be the most suitable option under certain conditions such as the following.

(1) The CLS does not require a very large sample, which would be the case when it is not required to produce estimates of the prevalence of child labour for many different regions, population groups, sectors of activity, or other types of domains.
(2) The CLS is a stand-alone survey, so that its sample can be designed as a compromise for meeting both types of information needs – of estimating the prevalence of child labour with necessary precision on the one hand, and of investigating the conditions and consequences of child labour with necessary detail on the other.
(3) Child labour is not too heterogeneous or extremely unevenly distributed for it to be 'captured' in a reasonable way by a general purpose sample of the population of children.
(4) The resulting compromise sample size is not too large for the in-depth investigation which the LCS component typically requires.
(5) And in any case, the resulting integrated interview is not too heavy to have an adverse effect on the quality (particularly completeness) of measuring the prevalence of child labour which is the concern of the CLS component. This is a common problem which has been encountered in many other types of surveys with similarly dual objectives.

When one or more of the above conditions are violated, it is necessary to at least consider the possibility of operationally separating the CLS and LCS components. In the case of such a separation, it would be generally appropriate to consider basing the LCS component on a subsample of the CLS. The objectives of the sub-sampling would be both to reduce the sample size for the LCS, and also to make it more concentrated and targeted to reflect the uneven geographical distribution of child labour. Specific subsampling procedures for this purpose are discussed in Section 7 below.

4.2.LCS linked to the CLS

An alternative to the integrated design is to conduct the CLS and LCS components as separate operations. However, these two cannot be stand-alone (i.e. entirely separated) surveys, but must be *linked* to each other in some way. The LCS sample can be identified on the basis of the CLS results in different ways, and this provides different forms of linkages between the two surveys. The best solution depends on the particular situation and objectives, but primarily on two factors: (i) the time gap between the two surveys; (ii) the quality of screening provided by the CLS in identifying the presence of working children.

The diversity of options in CLS-LCS linkage include the following.

A. In relation to selection of the sample within the CLS sample areas, we may take all or select a sample of:

(1) previously identified working children individually;
(2) households identified previously to contain a working child;

(3) households identified previously to contain any child in the age range of interest;
(4) all households interviewed in the previous sample;
(5) or possibly, all households selected in the previous sample (including non-respondents);
(6) all households listed in the CLS sample areas (i.e., obtaining a new LCS sample from existing CLS household lists); or
(7) households from re-listing of the CLS sample areas (i.e., updating the area lists before selecting a new sample).

B. In relation to 'eligibility for inclusion' of the CLS areas, we may take as eligible:

(8) CLS sample areas containing at least $x \geq 1$ working child(ren), or containing at least $y \geq 1$ household(s) with a working child;
(9) CLS sample areas containing at least $x \geq 1$ child(ren) or at least $y \geq 1$ household(s) with a child in the age range of interest; or
(10) all CLS sample areas.

C. In relation to the selection of areas, we may:

(11) take a sub-sample of the CLS sample areas;
(12) take all CLS sample areas; or
(13) select additional areas for LCS, linked to the CLS sample areas.

Some of the options in the three sets A–C can be combined, for instance: households identified previously to contain any child in the age range of interest (option 3), but only from a sub-sample of areas (option 11), those areas selected from CLS sample areas which contain at least one working child (option 8).

Note that among (1)–(7), all options except (1) require fresh identification of labouring children in the households included in the LCS sample. Option (1) provides the tightest link between the two surveys: the LCS sample is confined to the particular children identified to be engaged in child labour during the CLS. This makes the option almost the same as an integrated CLS-LCS design, except for the possibility of sub-sampling between the two operations because of their operational separation in time. With this option, any working children unidentified during CLS remain unidentified during the subsequent LCS operation. The opposite error – non-working children identified as working – is less important and in any case can be identified during the subsequent operation. With options (2)–(7), the LCS has a greater potential to refine the CLS estimates of child labour.

The categories from (13) to (1) represent increasingly close linkage between the two surveys. Going upwards from (13) to (1), especially from (7) to (1), the options generally become: more restrictive (the LCS sample is increasingly restricted to units identified in the CLS sample); more focussed (on the particular population of interest, i.e. labouring children); and more efficient exploiter of information already collected or the design already implemented in the CLS (hence more cost-effective, and less burdensome for the respondent).

But the options increasingly require: more information to be collected during the CLS and then fed-forward to the LCS (hence the operation becoming more costly and time consuming); and also more sensitive to changes over time (hence less suitable in the presence of a long time gaps between the two surveys).

Exclusion of certain CLS areas

Information on the reported level of child labour in the CLS is often useful in determining the cut-off level below which the areas may be altogether excluded from selection into the LCS. This is often necessary for practical reasons – it may simply not be cost-effective to attempt the LCS in areas containing none or very few reported working children. The cost considerations have to be balanced against the bias which such exclusion introduces.

Sub-sampling of CLS areas

Sub-sampling of CLS sample areas may be introduced in order to reduce the LCS sample size, and specifically to make the LCS sample more concentrated, i.e. confined to fewer sample areas with higher levels of child labour.

Expanding the original CLS areas

A brief comment will be useful on option (13). This option may be used when the LCS sample needs to include additional areas, beyond the CLS areas. This may, for instance, be because the CLS sample areas do not yield a sufficient number of sample cases (labouring children) for the LCS. Another motivation can be to enhance the LCS sample by selecting more cases from and around CLS areas found to contain concentrations of labouring children. An obvious way to expand the LCS sample is to include in it additional areas in the neighbourhood of CLS sample areas. Various techniques (such as 'adaptive cluster sampling', see Tompson and Seber, 1996) can be used for expanding the sample in this way, while retaining its probability nature.

5. Examples of diverse structures of child labour surveys

Tables 1 and 2 provide some essential information on the samples and the structure of linkages between different components for around 30 surveys on child labour. The tables show the diversity of the survey structure encountered. The survey-structural concepts have been explained in the preceding sections.

5.1. Linkage of CLS to a base survey

Table 1 classifies the type of linkage between the CLS and its base, which may be a larger survey (such as FLS) or the household listing operation. It can be seen that half the surveys (15 of the surveys reviewed) are *stand-alone surveys*, meaning that the survey is exclusively or primarily concerned with child labour.

The other common arrangement (13 of the surveys reviewed) is an *integrated survey*, involving the collection of the base survey and CLS information during the same operation. Here we distinguish modular versus combined surveys in the sense described in Section 3. The LFS forms the base for most of the combined surveys. It is also the case for several modular CLS's, but a variety of other types of surveys have also served as the base as shown in Table 1.

Another possibility, but a rare one, is to have a *linked survey*, where the child labour survey interview is operationally separated from that of the base survey, but the sample and some substantive information is fed-forward from the base survey.

The table also shows that most commonly CLS is a single round, one-time survey, though in one of six cases it has involved multiple rounds, typically four rounds corresponding to quarters of the year. A multi-round CLS is not affordable in most circumstances. Other information shown concerns sample size (n households) and its division into number of clusters (a) and sample-take per cluster (b=n/a). Sample size varies greatly, from 6,000 to 48,000 in the cases reviewed. Sample-take per cluster is even more variable – mostly in the range 5–50.

5.2. Linkage between CLS and LCS

Table 2 classifies the type of linkage between CLS and LCS components. It also indicates the substantive scope of the LCS – mainly whether it concerns primarily child labour (i.e. working children) or is a more inclusive survey of child activities or of children generally.

By far the predominant form has been an *integrated* CLS-LCS operation, in which information on prevalence of child labour and more detailed information on children who are found to be working are both collected during a single interview operation. A *linked* survey structure which permit a degree of operational separation between CLS and LCS has been used in only one-in-five of the surveys reviewed in the table.

Concerning the substantive scope of the LCS (irrespective of whether it is integrated or merely linked to the CLS), a majority are concerned primarily with economic activity of working children. However, almost one-half are broader in scope. A number are *child activity surveys* covering all types of activities of children including non-economic activities; some are even broader *children's surveys* covering in addition other areas such as children's health, housing and living conditions. This broader scope requires the coverage of a broader population – essentially of all children as in the case of the CLS.

Table 1. Child Labour Surveys: Base-to-CLS. Examples of diverse structures

Survey	Year	(1) Base-to-CLS	(2) Whether multi-round	(3) Sample size and clustering n=a*b	a	b
Azerbaijan	2006	Modular (LFS)		17,000	850	20
Portugal	1998	Stand-alone		25,000	1,150	22
Turkey	1994	Combined (LFS)		13,500		?
Turkey	1999	Combined (LFS)		20,000		?
Ukraine	1999	Combined (LFS)	4 rounds	48,000		?
Bangladesh	2002-2003	Stand-alone		40,000	1,000	40
Cambodia	1996	Modular (Socio- economic Survey)	2 rounds	9,000	750	12
Cambodia	2001	Stand-alone		12,000	6,000	20
Mangolia	2002-2003	Stand-alone (same sample as as LFS)	4 rounds	12,000	1,200	10
Nepal	1996	Modular (Migration and Employment Survey)		20,000	600	33
Pakistan	1996	Stand-alone *only a listing survey to identify target households*		140,000	1,860	75
Philippines	2001	Stand-alone		27,000	2,250	12
Sri Lanka	1999	Stand-alone	4 rounds	15,000	1,000	15
Ethiopia	2001	Stand-alone		44,000	1,250	35
Ghana	2001	Stand-alone		10,000	500	20
Kenya	1998-1999	Combined: LFS, Informal sector survey, child labour survey		13,000	1,100	12
Namibia	1999	Stand-alone *only a listing survey to identify target households*		8,000	270	30
Nigeria	1999	Stand-alone		22,000	2,200	10
South Africa	1999	Stand-alone		26,000	900	30
Tanzania	2000-2001	Modular (LFS)	4 rounds	11,000	220	50
Uganda	2000-2001	Modular (Demographic and Health Survey)		8,000	300	27
Zambia	2000	Modular (Multiple Indicator Survey)		8,000	360	22
Zimbabwe	1999	Linked (Indicator Monitoring (IM)-LFS) *IM-LFS provids lists of children in all sample households*		14,000	400	35
Belize	2001	Stand-alone		6,000	200	30
Costa Rica	2003	Combined (Multipurpose Household Survey)		11,000	?	?
Dominican Rep.	2000	Stand-alone		8,000	800	10
Honduras	2002	Modular (Permanent Multipurpose Survey)		9,000	1,800	5
Nicaragua	2000	Modular (ad-hoc LFS)		8,500	1,700	5
Panama	2000	Stand-alone		15,000	1,500	10
Georgia, Romania: similar to Ukraine						

An '?' indicates information not available from published survey report or other documentation.

***Source**: Compiled from national reports on surveys of child labour.*

Table 2. Child Labour Surveys: CLS-to-LCS. Examples of diverse structures

Survey	Year	1. Base-to-CLS	2. CLS-to-LCS	3. Sub-samling	4. LCS scope
Azerbaijan	2006	Modular	Linked	Sub-sample of areas and hhs with working children n=4.000 a=400 b=10 + special design for refugee children	Child labour (CL)
Portugal	1998	Stand-alone	Integrated		Chidren survey
Turkey	1994	Combined	Integrated		CL
Turkey	1999	Combined	Integrated		CL
Ukraine	1999	Combined	Integrated		Chidren survey
Bangladesh	2002-2003	Stand-alone	Integrated		CL"+employers" questionnaire
Cambodia	1996	Modular	Integrated		Children survey' + employers" questionnaire
Cambodia	2001	Stand-alone	Integrated		Chidren survey
Mongolia	2002-2003	Stand-alone	Integrated		Child activity survey
Nepal	1996	Modular	Integrated		CL
Pakistan	1996	Stand-alone	Linked	All households with labouring children (no subsampling) n=10.500 a=1.400 b=8	CL
Philippines	2001	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Sri Lanka	1999	Stand-alone	Integrated		Child activity survey
Ethiopia	2001	Stand-alone	Integrated		CL + schooling
Ghana	2001	Stand-alone	Integrated		Child activity survey
Kenya	1998-1999	Modular	Integrated		CL
Namibia	1999	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Nigeria	1999	Stand-alone	Integrated		Child activity survey'+ street children survey
South Africa	1999	Stand-alone	Linked	Sub-sample of hhs with a labouring child (all areas taken)	CL
Tanzania	2000-2001	Modular	Integrated		CL
Uganda	2000-2001	Modular	Integrated		CL
Zambia	2000	Modular	Integrated		CL
Zimbabwe	1999	Linked (IM-LFS)	Integrated		CL
Belize	2001	Stand-alone	Integrated		Chidren survey
Costa Rica	2003	Combined	integrated		Child activity survey
Dominican Rep.	2000	Stand-alone	Integrated		Child activity survey
Honduras	2002	Modular	Linked	2 in 5 subsample of household from all sample areas n=3.600 a=1800 b=2	CL
Nicaragua	2000	Modular	Integrated		CL
Panama	2000	Stand-alone	Integrated		Chidren survey
Georgia, Romania: similar to Ukraine					

***Source**: Compiled from national reports on surveys of child labour.*

6. Sample selection for a child labour survey (CLS)

This section describes some technical procedures for drawing the sample for a child labour survey (CLS) on the basis of the sample used for a larger survey of the general population, in particular the LFS.

6.1. The base survey (LFS)

In order to facilitate concrete discussion, we will assume the following design for the base survey. This is by far the most commonly used procedure for selecting population-based samples, especially in developing countries. It involves the selection of area units in one or more stages (often in only one stage) with probability proportional to a measure of population size of the area (p_i), and within each selected area, the selection of ultimate units with probability inversely proportional to the size measure. Below, summation Σ is over all areas in the population. Parameter 'a' refers to the number of areas (strictly 'ultimate area units') selected. If the current size of the area exactly equals the size measure p_i used for its selection, then f is the constant selection probability for any ultimate unit, parameter b is the constant number of units selected from any sample area, and $n = a \cdot b$ is the resulting sample size.

Selection probability for an area unit

$$f_{1i} = \left(\frac{a}{\Sigma p_i} \right) p_i = \frac{p_i}{I}, \text{ say} \tag{1}$$

Selection of ultimate units within selected area unit

$$f_{2i} = \left(\frac{b}{p_i} \right) \tag{2}$$

Overall selection probability of an ultimate unit

$$f_i = f_{1i} \cdot f_{2i} = \left(\frac{b}{I} \right) = f \text{, a constant.} \tag{3}$$

We will also assume the commonly used procedure of selecting area units systematically with probability proportional to size measure p_i from a list ordered in some meaningful way (PPS sampling). The systematic selection interval is $I = \Sigma p_i / a$, as defined above.

Dealing with very large and very small units

Special treatment is required in the selection of units of extreme (very large or very small) size. 'Very large' in the context of PPS sampling means a unit whose size measure exceeds the sampling interval, i.e. $p_i > I$. Such units may be segmented (divided into smaller areas) such that no segment exceeds I in size. (The segmentation may be applied to all units in the frame prior to sample selection, or only to the selected units using some objective rule not dependent on which particular units happen to be selected.) An alternative (and generally recommended) procedure is to treat large units as automatically selected. In this case, the selection equations become $f_{1i} = 1; \; f_{2i} = f$, the required overall constant rate. The sample size from the area is then proportional to its current size, $b_i = f \cdot p_i$.

'Very small' in the context of PPS sampling means a unit whose size measure is smaller than the required sample-take, i.e. $p_i < b$. Small units in the sampling frame may be grouped together (merged to form larger areas) such that no group is smaller than the required sample-take b. (As above, the grouping may be applied to all units in the frame prior to sample selection, or only to the selected units using some objective rule independent of which particular units happen to be selected.) Two commonly used alternative procedures are the following.

(1) Assigning small units a minimum size measure, $p_i = b$ for the area selection, and taking into the sample all final stage units in each selected area. The selection equations become $f_{1i} = (b/I) = f, f_{2i} = 1$. The original ultimate selection probabilities are retained unchanged, but, for a given sample size, the number of area units in the sample is increased. This can be inconvenient and expensive if there are too many small units in the frame.

(2) Excluding the smallest among small areas, with appropriate compensation. The following method has proved useful in a number of child labour surveys; it is particularly useful when the population includes many units, each of them with only a few (or even no) ultimate units of interest – as may well be the case in child labour surveys. In outline, the procedure is as follows. Let Σp_i be the total size measure of the set of 'very small' units to be sampled. The set of small units is ordered by unit size and divided into two parts. The first part is defined to consist of $A = \Sigma p_i / b$ *largest* units in this small-unit set. The size measure of each unit in this subset is increased to b, so that selection with interval I gives a sample of $a = A \cdot b/I = \Sigma p_i / I$ area units. All ultimate units in each selected area are retained in the sample. The second part consists of the remaining *smallest* of the small-unit set, with size measures $p_i < b_0$, say. For reasons of cost and practicality, these units are altogether excluded

from the sample, even though this is not properly a probability sampling procedure. It can be shown that a compensation for this exclusion is obtained by increasing the weight given to each selected unit in the first part by the factor (b/p_i), where p_i is the original size measure of the unit concerned.

6.2. Common structure but different design parameters between LFS and CLS

While the size measure in the LFS is the general population (or the population in the working ages), for a CLS it is an appropriately defined population of children exposed to the risk of child labour. It is generally the case that these two populations are closely related in size – the average difference between them being primarily a scaling factor. As noted, these similarities in the basic structure of the samples have important implications in the choice of appropriate survey structure for the CLS.

The procedure for selecting sample areas can also be the same in so far as the relevant base population sizes for the two surveys are nearly proportional to each other. In particular, if selection with probability proportional to size p_i is suitable for the LFS, it is reasonable to assume that, for practical purposes, the *same* size measures p_i as used in the LFS for PPS selection of areas are also appropriate for the same purpose in the CLS.

Despite similarity in the structure and distribution of the population to be sampled, the CLS and LFS will typically differ in a number of design requirements and parameters. In many countries, the LFS is a well established, regular or even a continuous survey. The CLS is more likely to be a new or recently instituted survey, conducted at best periodically and often only on an *ad hoc* basis. The resources available for the CLS are likely to be more limited, and often less certain. Increasingly, the LFS is required to produce separate estimates for different regions and sub-populations in the country, and produce these more frequently such as annually or even quarterly, while in most cases the primary objective of the CLS still has to be, at the first instance, the production of national-level estimates from time to time. In short, the LFS may be seen as a large, extensive and regular survey, and by comparison the CLS as a smaller, more intensive and less frequent survey. Consequently, the two types of surveys differ in relation to the choice of design parameters such as survey timing and frequency, sample size, the number of areas selected for the sample and the related sample size per area, allocation of the sample across different domains in view of different reporting requirements in the two surveys, the details of the stratification, and in the ultimate units for which data are collected.

6.3. Selection of CLS sample from a base survey such as the LFS

This subsection considers the procedure for selecting the sample for CLS from the sample for a base survey, typically the LFS.

One extremely important practical point should be noted at the outset when a sample is obtained by sub-sampling from an existing sample: full details must be recorded not only of the sub-sampling procedures, but must also be available for the existing sample used for sub-sampling. Unfortunately one finds examples of surveys in which details on the design of the existing sample were either not properly documented or had not been preserved. The most critical piece of information concerns probabilities of selection applied in the original sample. If such details are not available for an existing sample, it is desirable to look for alternative sources for sub-sampling.

Basic procedure

We begin with the simplest situations: for a given domain, a reduced number of areas are to be retained for the CLS from a given sample of LFS areas. Assume that the LFS is based on the commonly used PPS design described in Section 6.1. Let this sample contain ' a ' areas selected with probability proportional to a measure of area population size p_i. The objective is to select for the CLS a reduced number of areas, say $a' = g\,a,\ g < 1$, also with probability proportional to the *same* measure of size p_i. Given the common size measures, the procedure for sampling of areas from LFS to CLS is straightforward: select a sub-sample of LFS areas with a *constant probability* $g = a'/a$. Selection with a constant probability g can be achieved simply by applying to the LFS sample areas the *equal probability* systematic sampling procedure with interval $k = 1/g$. The result for the CLS is a PPS sample of areas with probability proportional to the same population size measure p_i. The selection equations for the LFS and the CLS areas are:

LFS: $f_1 = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}$, say, and CLS: $f_1' = \left(\frac{a'}{\Sigma p_i}\right).p_i = \left(\frac{a/k}{\Sigma p_i}\right).p_i = \left(\frac{p_i}{k\,I}\right)$,

where Σp_i is the sum of size measures for all areas in the population from which the LFS sample was selected.

This simple procedure involving subsampling of areas at a constant rate applies to any arbitrary choice of size measures p_i provided that they are the same in the two samples.

Dealing with 'very large' areas in sub-sampling

Additional steps are involved in the sub-sampling procedure in dealing with areas in the original sample which were selected using special procedures because

they were considered to be too 'small' or too 'large' for the normal PPS procedure, in the sense described earlier.

We first consider very large units, i.e. units with $p_i > I$. Assume that in the LFS such areas have been selected with $f_{1i} = 1; \quad f_{2i} = f$ (see Section 6.1). For the purpose of selecting a $(1/k)$ sub-sample of these 'self-representing' LFS areas for the CLS, two groups among these need to be distinguished.

Group 1: $p_i > (k\,I)$. These are the largest units. All of these units must be retained in the CLS with probability =1, as in the LFS. At the final stage, ultimate units can be selected with the overall selection rate, say f', required for the CLS.

Group 2: $I \le p_i \le (k\,I)$. All these large units do not get selected into the CLS automatically, though that was the case in the LFS. For these units, a proper PPS sample of areas can be selected for the CLS with $f'_{1i} = (p_i / k\,I) \le 1$. For a self-weighting design with overall sampling rate f', the final stage selection probabilities would be $f'_{2i} = (f'/f'_{1i})$.

Dealing with 'very small' areas in sub-sampling

The sub-sampling procedure becomes more complex when dealing with very small units. In a two-stage design, the procedure for handling very small areas in LFS-to-CLS sub-sampling depends on the details of the LFS sampling at both stages. For describing the sub-sampling procedure, we assume the LFS sample selected according to equations (1)–(3).

For the CLS, sub-sampling from the LFS involves two steps: selection of 1 in k sample areas, $a' = (a/k)$, and then (an average or target) selection of an expected number b' of ultimate units per area included in the CLS, giving $n' = a'\,b'$ as the CLS target sample size. Note that the two (LFS and CLS) samples may not necessarily overlap as concerns the ultimate units, though often they do and $b' \le b$. Also, both for the LFS and the CLS, these parameters may differ from one sampling domain or stratum to another. It is sufficient here to describe the procedure for one such domain.

The full selection equations for self-weighting CLS design, corresponding to (1)–(3), are:

Selection probability for an area unit

$$f'_{1i} = \left(\frac{a'}{\Sigma p_i}\right) p_i = \frac{p_i}{k\,I} = \frac{p_i}{I'}, \text{ say} \qquad (4)$$

Selection of ultimate units within selected area

$$f'_{2i} = \left(\frac{b'}{p_i}\right) \qquad (5)$$

Overall selection probability of an ultimate unit

$$f_i' = f_{1i}'\ f_{2i}' = \left(\frac{b'}{I'}\right) = \left(\frac{1}{k}\,\frac{b'}{b}\right) f = f', \text{ a constant.} \qquad (6)$$

The procedure for selecting for the CLS a sub-sample of very small LFS areas, i.e. areas with size measures $p_i < b$, depends on how these areas were selected into the LFS sample itself. In addition, we have also to consider the relationship of the size measure p_i to the required sample-take b' in the CLS. Table 3 shows the selection equations for very small areas in CLS, corresponding to the two procedures for their treatment in the LFS described in Section 6.1, namely:

Procedure (1). Assigning small units a minimum size measure, $p_i = b$ for the area selection, and taking into the sample all final units in each selected area.

Procedure (2). Excluding smallest of small areas, with appropriate compensation.

Table 3. Sub-sampling of LFS areas for the CLS

	Condition	LFS		CLS		$\frac{f_1'}{f_1}$
		f_1	f_2	f_1'	f_2'	
	Very large areas					
1	$p_i > (k\ I)$	1	b/I	1	$b'/k\ I$	1
2	$I \le p_i \le (k\ I)$	1	b/I	$p_i/k\ I$	b'/p_i	$p_i/k\ I$
	Normal areas (majority of the areas)					
3	$b, b' \le p_i < I$	p_i/I	b/p_i	$p_i/k\ I$	b'/p_i	$1/k$
	Very small areas – area selected in LFS using Procedure (1)					
4 *	$b \le p_i < b'$	p_i/I	b/p_i	$b'/k\ I$	1	$b'/k\ p_i$
5	$b' \le p_i < b$	b/I	1	$p_i/k\ I$	b'/p_i	$p_i/k\ b$
6	$p_i \le b, b'$	b/I	1	$b'/k\ I$	1	$b'/k\ b$
	Very small areas – difference from above if selected in LFS using Procedure (2)					
5	$b_0, b' \le p_i < b$	as in case 5 above				$1/k$
6	$b_0 \le p_i \le b, b'$	as in case 6 above				$b'/k\ p_i$
7	$p_i < b_0$	areas not included in LFS or CLS sample				--

** Note that only one of the two cases, 4 and 5, can apply in any particular situation. See Section 6.1 last paragraph, for definition of b_0: units with size smaller than this limit were dropped from the LFS.*

6.4. Sample allocation and reporting domains

Apart from differences in the required sample size and clustering, the CLS may differ from the 'parent' LFS also in the requirements concerning sample allocation and stratification.

For instance, the LFS may be allocated disproportionately for the purpose of producing sub-national estimates with over-sampling of small regions or other reporting domains, while this may not be required for a CLS when it is based on a smaller sample aimed primarily at producing national-level estimates, or producing breakdowns only for a few major domains.

In any case, we can generally expect the CLS to have a smaller number of reporting domains than a bigger survey like the LFS; furthermore, the sampling rates in the CLS sample are often more uniform. Typically, the reporting domains for the CLS would be groupings of the LFS reporting domains – e.g. major regions of the country rather than individual provinces or districts. It is unlikely to have the situation in which the more detailed LFS domains cut across boundaries of the more aggregated CLS domains.

Significant differences in the required stratification are unlikely. Certain common stratification criteria, such as geographic location and degree of urbanisation, are commonly used in all types of household surveys. Often these are practically the only stratification criteria available. Where available, additional useful criteria (ethnicity, predominating occupation, literacy rate, mean level of income, etc., for the sample area) also tend to be similar for different social surveys. The CLS and LFS are likely to be even more similar in terms of stratification because of their shared or similar subject matter. Generally, differences in stratification requirements are likely to arise only in relation to differing requirements between the two surveys in terms of sample allocation.

7. Sample selection for labouring children survey (LCS)

7.1. The required sample structure

As noted, a critical issue of practical importance is whether the distinct CLS and LCS objectives can be satisfactorily met through a single integrated survey, or it is better to organise them as two separate – but nevertheless linked – operations. In the latter case, a related question is whether the two components can be based on the same sample of units, or the LCS should be a subsample of the CLS, smaller in size and possibly also with a different structure.

The main difference between the CLS and LCS concerning the sample structure is that in the former the primary focus was on the measurement of prevalence of child labour among the population of all children, and consequently that population formed the base for the design and selection of the CLS sample. The base for the LCS is the population of working children. Consequently, the

selection of LCS sample areas requires information on the number of labouring children in each area in the 'frame' from which the survey areas are to be selected. Such information is not normally available in general-purpose population-based frames. It is this reason that makes it necessary to select the LCS as a sub-sample of areas for which such information has been collected, such as sample areas from the larger CLS.

This section considers the procedure for selecting the sample for LCS from the sample for CLS as the base. For the development and exposition of the CLS-to-LCS sampling procedure, we will assume the basic sampling scheme for the CLS to be equations (4)–(6); actually for notational simplicity, we will use their equivalent, equations (1)–(3).

In a LCS, where the base population of interest is labouring children, an appropriate design will involve the selection of area units with *probability proportional to the number of labouring children* (c_i) in the area, and then the selection, within each selected area, of such children with probability inversely proportional to c_i:

Selection probability for an area unit

$$f'_{1i} = \left(\frac{a'}{\Sigma c_i}\right) c_i = \frac{c_i}{I'}, \text{ say} \tag{7}$$

Selection of ultimate units within selected area

$$f'_{2i} = \left(\frac{b'}{c_i}\right) \tag{8}$$

Overall selection probability of an ultimate unit

$$f'_i = f'_{1i}\ f'_{2i} = \left(\frac{b'}{I'}\right) = f', \text{ a constant.} \tag{9}$$

This design differs from that of the CLS discussed earlier in a number of respects.

(1) The LCS design depends on the categories of children included in the target population – for example the definition of child work or labour used, or the specific types of child activities included. The measure of size c_i refers to the numbers of children in the categories of interest. These values cannot be assumed known for all areas in the population. It is assumed that these are obtained or estimated from a base survey (LFS, or more appropriately a CLS), but only for the areas enumerated in that survey. Hence the LCS must be confined to a sub-sample of areas in the first survey. Within common sample areas, the samples of ultimate units may of course be different or overlapping.

(2) The base population of labouring children is likely to be much more unevenly distributed over sample areas compared to the general population of children. A few areas may contain high concentrations, and many areas only very low numbers of working children.
(3) In particular, there may be many 'zeros', i.e. areas containing no labouring children of interest in the LCS. Problems such as the presence of extreme ('very large' or 'very small') areas, defined here in terms of the number of working children the area contains, are likely to be much more widespread than those in the CLS.
(4) The sample size of the LCS is likely to be (or at least should be in a good quality survey) much smaller because of its intensive nature.

7.2. Selection of area units

Given a sample of areas selected according to equations (1)–(3) in the base survey, how to obtain a sample of areas of the type described by (7)–(9) for the LCS? This can be achieved by selecting a sub-sample of areas from the first sample with PPS, the area measures of size for this sub-sampling being the ratio (c_i/p_i). This can be expressed as:

$$g_i = a'\left(\frac{(c_i/p_i)}{\Sigma_s(c_i/p_i)}\right),\ f'_{1i} = g_i \cdot f_{1i} \qquad (10)$$

where the sum is taken over all areas in the base (CLS or LFS) sample, as indicated by the subscript $'s'$; g_i is the probability of area i from the base sample being selected into the LCS; and a' is the number of areas to be selected for LCS. This with (1) gives:

$$f'_{1i} = g_i \cdot f_{1i} = a'\left(\frac{(c_i/p_i)}{\Sigma_s(c_i/p_i)}\right)\left(\frac{a}{\Sigma p_i}\right)p_i = \left(\frac{a'}{k_s}\right).\frac{c_i}{\Sigma c_i},$$

$$\text{with } k_s = \frac{\Sigma_s(c_i/p_i)/a}{\Sigma c_i/\Sigma p_i} \qquad (11)$$

Hence sub-sampling procedure (10) results in a sample of areas selected with probabilities proportional to size measure c_i, as required. In the above equation, Σ is the sum over areas in the population, Σ_s is sum over areas in the base sample, and k_s is a constant determined by the population and base sample characteristics, independent of the LCS sample or particular area i.

Note that if the LCS sample of a' cluster were selected directly from the population, with area selection probability proportional to the size measure c_i (a

function of the number of labouring children in the area), the selection equation would have been (7) instead of (11). Thus k_s is a factor surmising the effect of selecting this sample 'indirectly', via the base survey. If c_i is strictly proportionate to p_i in all areas, it can be seen that $k_s = 1$. In fact factor k_s is not known since c_i values are not known for *all* areas in the population. It also depends on the particular sample which happens to be selected in the first survey – hence (11) does not provide the true selection probabilities in the sense of expected values over all possible base samples. Nevertheless, the value of this factor is expected to be close to 1.0, since its numerator and denominator both estimate average of the ratio (c/p): the numerator is the average in the base sample of separate ratios (c_i/p_i) while the denominator is the combined ratio $\Sigma c_i/\Sigma p_i$ of the same quantities in the population. In any case, this factor does not affect the *relative probabilities* of the area units selected into the final sample, since the factor is the same for all these units. The units are therefore selected with relative probabilities proportional to their size measures c_i.

7.3. Dealing with 'very large' and 'very small' areas

Units with extreme characteristics are likely to occur in the LCS more often than in surveys like the LFS or CLS. Care is required to ensure that correct selection probabilities are achieved for such units. Such problems concerning unit size of course also occur in selecting the base sample. However, these aspects for the base sample generally do not make the treatment of such cases in the LCS more complicated. This is because different sets of areas are involved as 'extreme' cases in the base and the LCS samples since the two use different types of size measures. Hence, generally the two sets can be dealt with separately.

Very large areas

In the context of sub-sampling from the base sample to obtain a sample of areas for the LCS, 'very large' actually refers not to the area population size but to the *degree of concentration of child labour* in it, i.e. to very high values of the ratio (c_i/p_i). This is because this ratio is used as the size measure in (10). Very large are units for which this measure exceeds the sampling interval used in the selection of LCS areas from the base sample:

$$(c_i/p_i) \geq I_s \text{ where } I_s = \frac{1}{a'}\Sigma_s(c_i/p_i).$$

These areas can be treated in the same way as in other cases described in previous sections. For instance, one may redefine any measure of size $(c_i/p_i) \geq I_s$ as $= I_s$, so that any such area in the first sample is taken into the

LCS sample with certainty. These areas thus retain the original probabilities of selection into the base sample unchanged for the LCS. As to the ultimate stage of selecting households or persons in these sample areas, the ultimate stage sampling rate f'_{2i} in equation (8) may be correspondingly adjusted so as to keep the required overall selection probability f' unchanged.

Very small areas

For large areas the PPS sampling procedure needed adjustment only because the size measure (c_i/p_i) exceeded the sampling interval I_s. Therefore areas were identified as being large or not large on the bases of proportion (c_i/p_i) of working children among all children (or persons) in the area. By contrast, areas are defined as being 'very small' in terms of the number of ultimate units they possess or contribute to the sample, i.e. the expected absolute number of working children c_i in the area.

The presence of small c_i values has important practical consequences.

(1) First of all, it should be emphasised that in the design described above of selecting areas from the base survey with probability proportional to (c_i/p_i), areas for which no working children have been reported in the base survey, $(c_i = 0)$, are automatically excluded from the LCS sample. Formally, this of course is also true of areas with no population $(p_i = 0)$ for the selection of the sample areas in the base sample with probability proportional to p_i. But in practice the two situations are quite different. Areas with no population $(p_i = 0)$ tend to be rare and of no interest to the survey in any case, but areas with no working children $(c_i = 0)$ may be very common. Furthermore, the situation with respect to the later (working children) is likely to be much more changing compared to the situation with respect to the former (population). Hence the information on the presence or otherwise of working children in an area needs to be quite fresh.

(2) Even when not exactly zero, very small values of c_i are much more likely to occur than small p_i values. This is because of the uneven distribution of child labour across sample areas. The practical question arises as to whether a lower bound should be put for automatic exclusion from the sample of areas with c_i values below that limit.

With 'very small' areas defined in terms of the size measure c_i, small are the areas whose size measure is smaller than the required sample-take at the ultimate stage, i.e. $c_i < b'$ in equation (8). Procedures for dealing with very small areas are similar to those described earlier. However, different procedures may be used

for the selection of households within sample areas; these procedures are probably more varied in labouring children surveys, compared to other, more general surveys of the population.

7.4. Obtaining sufficient sample size for the LCS

Take-all sampling

An obvious concern in LCS is to ensure that the required sample size of working children can be achieved in practice. This can be a problem if the prevalence of child labour is lower than what was assumed at the time of sample design, or if because of poor quality the previous survey missed a large proportion of units subject to child labour. When such problems exist, a 'compact cluster' or 'take-all' design can be an interesting option. In this design all relevant units (e.g. households with a working child) in a selected area are taken into the sample, possibly with an upper limit on the maximum number to be selected. Selection probabilities of ultimate units, as well as sample takes per area, will generally vary in such a design.

Expanding the size of first sample areas

It can happen that the type of areas originally selected in the base sample are generally too small to yield the required number of cases for the LCS. In such situations it may be necessary to consider whether some of the areas – perhaps those with high concentrations of labouring children, areas which are also likely to have such high concentrations in the neighbourhood – can be expanded in physical size to include additional neighbouring areas. One simple procedure is to group areas in the frame into exhaustive and non-overlapping larger areas. Each original smaller area selected brings into the sample the whole of the larger area it belongs to. In statistical terms, the resulting sample would be essentially equivalent to the selection of the larger units, with the probability of selection of such a unit equalling the sum of the selection probabilities of all the smaller units contained within it. Adaptive cluster sampling is another, more sophisticated approach which may be useful and feasible in certain circumstance.

REFERENCES

HUSSMANNS, R, MEHRAN, F. and VERMA, V., 1990. *Surveys of Economically Active Population, Employment, Unemployment and Underemployment: An ILO Manual on Concepts and Methods*. Geneva: International Labour Organisation.

INTERNATIONAL LABOUR ORGANISATION, 2004. *Child Labour Statistics: Manual on Methodologies for Data Collection through Surveys*. Geneva:

International Labour Organisation, International Programme for the Elimination of Child Labour (IPEC).

KORDOS, J., 2008. Bookreview: Sampling for Household-based Surveys of Child Labour, by Vijay Verma*, Statistics in Transition, 9(3), 587–590.*

PORTUGAL MINISTRY OF LABOUR AND SOLIDARITY, 1998. *Child labour in Portugal: Social Characterisation of School Age Children and Their Families*. Lisbon: MTS.

TOMPSON, S.K. and SEBER, G.A.F., 1996. *Adaptive Sampling*. John Wiley & Sons.

VERMA, V., 2008. *Sampling for Household-based Surveys of Child Labour*. Geneva: International Labour Organisation.