Resources for Quantitative Surveys on Child Labor

SEPTEMBER 30, 2018

Sarah Liuzzi, Shira Mitchell, Paolo Abarcar, Abbie Turiansky, and Sarah Dolfin

Background

As part of its mission to combat child labor, the Office of Child Labor, Forced Labor, and Human Trafficking (OCFT) in the U.S. Department of Labor (DOL) provides grants to support research that can help the international community to better understand and prevent child labor. DOL's Chief Evaluation Office contracted with Mathematica Policy Research to review guidance provided to and reports drafted by OCFT grantees to: (1) identify key research challenges and areas of opportunity to support grantees' continuous improvement, and (2) develop resources to support researchers in the area of child labor. This document presents resources for conducting child labor research based on existing research and best practices.

M50 MATHEMATICA Policy Research

Resources for Quantitative Surveys on Child Labor

SEPTEMBER 30, 2018

Sarah Liuzzi, Shira Mitchell, Paolo Abarcar, Abbie Turiansky, and Sarah Dolfin

Submitted to:
U.S. Department of Labor
200 Constitution Avenue, NW
Washington, DC, 20210
Project Officer: Janet Javar
Contract Number: DOLQ129633249/DOL-OPS-15-U-00183

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Sarah Dolfin
Reference Number: 50174

Disclaimer

This report was prepared for the U.S. Department of Labor (DOL), Chief Evaluation Office by Mathematica Policy Research, under contract number DOLQ129633249. The views expressed are those of the authors and should not be attributed to DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

MATHEMATICA Policy Research

ACKNOWLEDGMENTS

The authors acknowledge several people who contributed to and supported the research. At the U.S. Department of Labor, Janet Javar and Kuangchi Chang in the Chief Evaluation Office and Kevin Hong, Lauren Damme, and Celeste Lemrow at the Office of Child Labor, Forced Labor, and Human Trafficking, Bureau of International Labor Affairs provided valuable direction, feedback, and materials that were instrumental in developing the report. Ana Dammert at Carleton University, Andrew Dillon at Michigan State University, Deborah Levison at the University of Minnesota, and Margaret Meagher at African Leadership Academy provided expert feedback and insightful suggestions as members of the Technical Working Group for the project. Sarah Hughes and Ravi Goyal at Mathematica reviewed the report and provided comments that greatly enhanced its quality. Katie Naeve and Alexander Johann at Mathematica made important contributions to the analysis of grantee work. Rose Sullivan, Colleen Fitts, Sheena Flowers, and Shantal Alston helped design and prepare the report and Sharon Richards and Betty Teller provided editorial assistance. Carmen Ferro and Rich Clement led the production of the video resources developed as part of the project to complement the resources in this report. Kevin Fleury of the Creative Circle designed and implemented the animation for the animated videos.

Contents

I. INTRODUCTION 1

II. SAMPLING APPROACHES FOR ESTIMATING THE PREVALENCE OF CHILD LABOR 7

Exhibit II.1. Sampling Worksheet 23

III. SURVEY DESIGN AND DATA COLLECTION 29

Exhibit III.1. Checklist for survey design and data collection 40

Exhibit III.2. Checklist for defining and measuring child labor 43

Exhibit III.3. Annotated sample household survey on child labor 45

IV. DATA FORMAT AND VALIDITY 62

Exhibit IV.1. Checklist for data format and validity 68

V. ANALYSIS AND REPORTING 70

Exhibit V.1. Checklist for clear reporting 82

VI. PREPARING PUBLIC USE DATA FILES 85

Exhibit VI.1. Checklist for public use data files 99

VII. ONLINE RESOURCES, BY CHAPTER 101

TABLES

I.1. Strategies to strengthen child labor survey research ... 3
I.2. Summary of resources and suggestions for using them ... 5
III.1. Challenging and easier questions for young child respondents ... 36
IV.1. Example of household and child data set structures ... 65
IV.2. Example of variables with variable labels, response values, and response value labels ... 65
V.1. Example: Estimated prevalence of child labor in project regions ... 73
V.2. Example: Estimated prevalence of child labor in project regions at baseline and follow-up ... 74
VI.1. Linking files provide a link between PII and public use files ... 90
VI.2. Data elements with very few observations ... 91
VI.3. Top-coding continuous variables above the 95th percentile ... 94
VI.4. Aggregating numeric variables with very few observations ... 95
VI.5. Combining categories for variables with very few observations ... 96

FIGURES

II.1. Steps to estimate prevalence of child labor 8
II.2. Households sampled from the population of interest 9
II.3. Households sampled by simple random sampling (SRS) 11
II.4. Households sampled by one-stage cluster sampling 12
II.5. Households sampled by two-stage cluster sampling 12
II.6. Households sampled by one-stage cluster sampling, not stratified on district 13
II.7. Households sampled by one-stage cluster sampling, stratified on district 13
II.8. Households ordered by household size, within a sampled PSU 14
II.9. Households sampled by systematic sampling (after being ordered by size) 15
II.10. Households sampled by simple random sampling (SRS) 15
II.11. The simplest sampling frame 16
II.12. A cluster sampling frame 17
II.13. Stage I of the basic sampling design 18
II.14. Stage II of the basic sampling design 19
III.1. Visual aid for time-related questions for young children 37
V.1. Example that is overloaded with information (Do not follow this example) 76
V.2. Example that presents the right amount of information (Follow this example) 76

I. INTRODUCTION

As part of its mission to combat child labor, the U.S. Department of Labor's Office of Child Labor, Forced Labor, and Human Trafficking (OCFT) provides grants to support the work of organizations around the world that implement projects to keep children out of the child labor (referred to as OCFT grantees). Some OCFT grantees gather data to estimate, both before and after project implementation, the prevalence of child labor in the areas that they serve. They use the data to produce baseline and follow-up reports to document changes in child labor from before they started their activities to after they completed them.

In 2017, Mathematica conducted a detailed review of the work of a purposively selected sample of nine OCFT grantees from nine countries where grants were awarded from 2012 to 2015. The objective of this review was two-fold: to identify areas of opportunity to support the continuous improvement of data collection and analysis and to develop resources that can help researchers to do so. These grantees are a small part of the broader research community working on child labor issues, yet they face many challenges that the larger research community shares. Based on the review of grantees' work, and informed by other available best practices, this report presents a set of resources for conducting child labor research, which cover the following topics:

- Sampling design
- Survey development
- Data format and validity
- Analysis and reporting
- Preparation of public use data files

In Section A of this chapter, we briefly summarize some key research challenges that the research community—including OCFT grantees—face and options to improve data collection and analysis. In Section B, we describe the resources provided in Chapters II–VI of this report by topic and resources that accompany it and suggest how researchers can best maximize their use.

A. Overview of options for overcoming key challenges in quantitative child labor survey research

Researchers face numerous challenges when they implement quantitative surveys to estimate the prevalence of child labor and issues associated with it. Researchers must draw a sample that is representative of the population of interest; doing so requires acquiring a household listing or creating one. The definition of *child labor* used must be consistent with research objectives, and surveys must include all components of the child labor definition. As with all survey research, researchers must take care to verify the quality of the data collection process from survey design and training to data entry and cleaning. Report authors may have to distill and convey key findings drawn from large quantities of survey data. If data are to be made publicly available, researchers must take steps to protect respondents' confidentiality.

Particular challenges facing child labor researchers are those of gathering sensitive data and data from children. Researchers must gather data either from caregivers, who may not be aware of children's working conditions, or from children, who may struggle to understand research questions crucial to measuring child labor on topics such as work schedules, weights lifted, or working conditions—especially if they are young.

In response to the challenges child labor researchers face, this document outlines strategies researchers may wish to take to confront those challenges at each major step in the research process. Table I.1, organized to align with the chapters of the report, summarizes those strategies. Chapters II through VI provide more detail on these and other strategies. Each chapter includes written text, and several chapters include additional written resources, such as worksheets and checklists. This document is also accompanied by videos and Excel templates. Section B of this chapter describes the full set of resources and how to use them.

Table I.1. Strategies to strengthen child labor survey research

Sampling
• When possible, use templates, such as those that accompany this report, to guide key decisions related to the sampling approach and analysis plans. • Document key elements of the sampling procedure, including the sampling design and choice of sampling frame, choice of design effect, and inferential concerns (weights, nonresponse, and statistical uncertainty) in reports or methodological appendices. • For those collecting both baseline and endline data, avoid modification of survey instruments or data collection procedures between baseline and endline, as modification might threaten comparability.
Survey Development
• Use existing tools, such as International Labour Organization (ILO) survey instruments or national surveys, tailored to specific research goals and context. • Pilot surveys and revise instruments and data collection procedures based on pilot results. • Use a survey development checklist, such as the checklist provided in this report, to avoid missing key steps. • In the survey, state the definitions of *household, household head,* and *targeted respondents* clearly. • Take care when writing survey questions for children; be sure that children of the full age range to be surveyed will understand the questions. • Assess the survey instrument to be sure its questions capture age data accurately, as some respondents may not know their date of birth. • Document the translation process for surveys either in a report or in an accompanying methodology document.

Table 1.1, cont.: Strategies to strengthen child labor research

Data Format and Validity

- Use labels and codes that are consistent with surveys when entering data.
- Code categorical data in a way that is conducive to analysis. For example, format categorical variables as numeric in data sets and provide labels.
- Check for issues underlying missing data, outliers, and impossible and implausible values.
- Check and address instances of noncompliance with skip patterns (ideally during piloting or early in the data collection process).
- Clean variables with impossible or implausible values. Confirm that errors are addressed in cleaning.

Analysis and Reporting

- In reports, include all essential sections, following a structure similar to the structure outlined in this report.
- Take care to avoid common pitfalls associated with the presentation of results, such as unclear figure labels or presentation of too much information in a single figure.
- Use standard errors, coefficients of variation, or 95 percent confidence intervals for the estimates when comparing statistics.
- Include central results in the main text and less-central tables and figures into an appendix.
- Conduct in-depth analysis to explore what correlates with child labor. Discuss unexpected findings and any qualitative results.
- Reflect on how lessons learned may influence implementation of interventions to combat child labor.
- Assess whether data depend on which enumerator administered the survey (enumerator effects).

Preparation of Public Use Data

- Determine necessity of preparing data sets for public use, restricted use, or both.
- Follow steps for creating public use data sets based on handling of geographic and interviewer identifiers, top- and bottom-coding, removal of rare values, and restriction of household and personal information. Remove respondents' names and exact dates of birth from their data sets.
- Prepare "user guides" to accompany public use files. These guides would orient the user to how to use the data files and would describe changes made to the data during the cleaning process.

B. Resources in and accompanying this report and how to use them

These resources cover the full research design and implementation process for implementing quantitative child labor surveys but emphasize the areas we believe are the most likely to be challenging for child labor researchers. Table I.2 provides a summary of the resources that are included in or accompany this document and suggestions for how to use them.

Table I.2. Summary of resources and suggestions for using them

Chapter II. Sampling design (p. 7)
The sampling resources include (1) a chapter on the principles of sampling and the steps to carrying out a basic sampling approach (p. 7); (2) a worksheet to guide a user through the approach (p. 23); (3) an Excel template to carry out sampling; and (4) a set of videos that walk the user through the resources. Because sampling is technical, allow ample time to review the text and resources. Determine your sampling plan before committing to data collection or procuring data collection services. Before beginning data collection or data collection procurement, use the resources on sampling to understand principles of sampling and to implement your own sampling plan. You may follow these steps to use the resources to develop your sampling plan: **1.** Watch Video 1: Introduction **2.** Watch Video 2: Define the population of interest **3.** Read Chapter II on sampling (p. 7) **4.** Complete the sampling worksheet (p. 23) **5.** Watch Video 3: Develop a sampling design and sampling frame **6.** Watch Video 4: Choose sample sizes with the template **7.** Complete the Excel template for choosing sampling sizes Later, once you have collected and cleaned your data, follow steps outlined in Chapter V on analysis and reporting to estimate child labor prevalence and the standard errors and confidence intervals for those estimates.
Chapter III. Survey development (p. 29)
The chapter on survey development covers designing survey instruments, implementing them in data collection, and defining and identifying child labor (a critical component that drives decisions on whom to interview and what questions to ask). Resources on this topic include a checklist for survey design and data collection (p. 40), a checklist for defining and measuring child labor (p. 43), and an annotated survey (p. 45). To make the most of this chapter, review these resources before designing the survey and revisit them, as needed, as survey instruments are developed. **1.** Read Chapter III on survey development (p. 29) **2.** Review the checklist for survey design and data collection (p. 40) **3.** Review the checklist for defining and measuring child labor (p. 43) **4.** Review the annotated household survey on child labor, which provides examples to follow and examples of thing to watch for when customizing surveys (p. 45)

Chapter VI. Data format and validity (p. 62)

The data format and validity chapter discusses how to code and format data gathered from the surveys in a way that preserves their quality for analysis.

1. Read Chapter IV on data format and validity before developing an approach to data entry or data cleaning or completing survey development (p. 62)
2. Review the checklist for data format and validity (p. 68)

Chapter V. Analysis and reporting (p. 70)

The analysis and reporting chapter covers the writing of baseline and follow-up reports following data collection.

1. Read Chapter V on analysis and reporting before outlining or writing your report (p. 70). This chapter includes templates for presenting results, descriptions of characteristics of clear and complete tables and figures, and statistical analysis required to report estimates of child labor prevalence.
2. Review the checklist for clear reporting (p. 82).

Chapter VI. Preparation of public use data (p. 85)

For those researchers who will make their data available to the public, Chapter VI on preparing public use data (p. 85) describes steps to ensure respondents' confidentiality while maximizing the utility of the data for eventual users.

1. Early in the research process, read Chapter VI on the preparation of public use data (p. 85). Do this before budgeting to be sure that you allow sufficient time and resources for the complex process of creating public use data sets, if necessary.
2. Review the checklist for public use data files (p. 99).

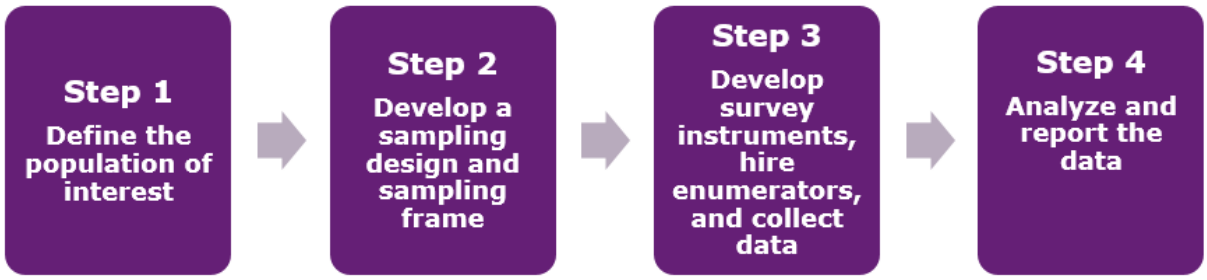
Chapter VII. Other online resources (p. 101)

Researchers face diverse challenges when developing and implementing a plan to gather data to estimate the prevalence of child labor. These resources will not cover every challenge that might arise, and some researchers may wish to review more detailed resources on topics covered here. The annotated list of online resources provided in this chapter is a compilation of in-depth resources on these topics. All resources listed are freely available online. Consult these resources as needed at any stage in the research process.

II. SAMPLING APPROACHES FOR ESTIMATING THE PREVALENCE OF CHILD LABOR

To estimate the prevalence of child labor in a geographic target area from survey data, follow the four steps shown in Figure II.1.

Figure II.1. Steps to estimate prevalence of child labor



The sections below provide descriptions of each of the four steps, with definitions (in boldface text) of associated concepts. Some of these concepts are quite technical; consider consulting a statistician for assistance.

This chapter describes Steps 1 and 2. The Sampling Worksheet that follows includes questions to work through for these first two steps. Other chapters cover Steps 3 and 4. For more information on Steps 1, 2, and 4, we direct researchers to sections from the ILO manual, *Sampling for Household-Based Surveys of Child Labour*. This is a comprehensive resource on sampling for household-based surveys of child labor, and its content is consistent with the steps outlined in this chapter. We refer to specific sections of the manual in parentheses and red text throughout this section. A set of five videos that accompany this document cover steps 1, 2, and 4. We recommend viewing these videos in addition to reading this chapter.

A. Step 1: Define the population of interest

The population of interest (ILO Section 3.3.1, the "survey population") is the population you want to study. You will study it by collecting survey data on some households and children in this population (a "sample"); see Step 2. We use the term "population" to mean the population of interest. This population is defined by answering three questions:

- What is the geographic target area (for example, Districts A, B, and C) you want to study?
- What is the time period you want to study? (For example, a project starting April 2018 may measure baseline outcomes from January to March 2018. For some projects, you may also want to cover certain agricultural seasons or times when school is in session.)
- Who are the people (for example, children ages 5 through 17) you want to study?

In this chapter, we assumes that the population of interest includes all children ages 5 through 17 in a project's geographic target area.

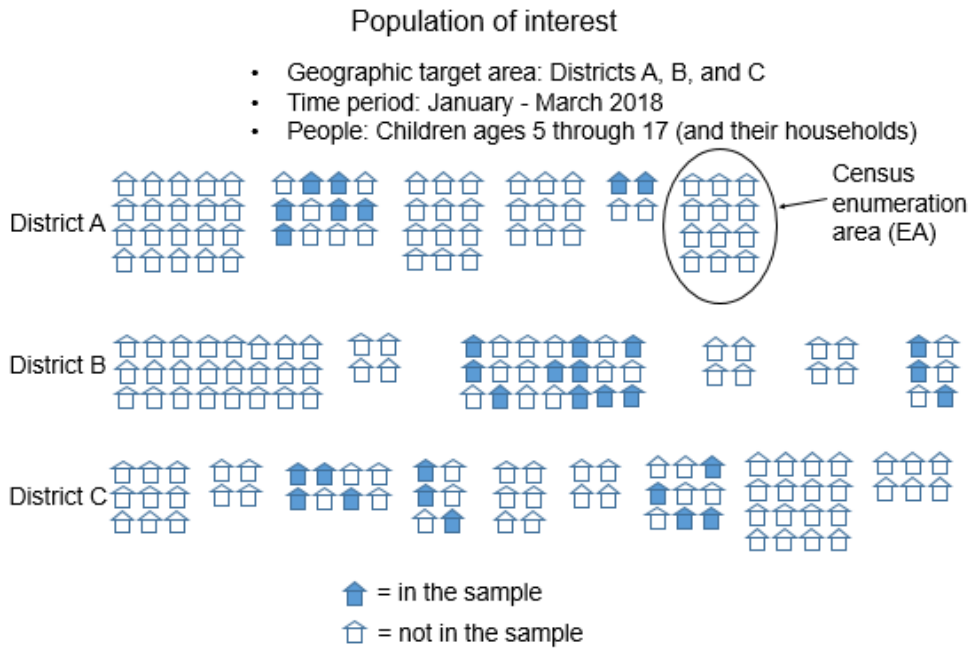
B. Step 2: Develop a sampling design and sampling frame

Suppose you want to know the proportion of the population in child labor. It is too expensive to survey everyone, so you survey a **sample**, which is a subset of the population—some households and their children. You use survey data from the sample to calculate an **estimate** of the proportion of the population in child labor. The **standard error** measures how close you expect the estimate from the sample to be to the population proportion. Small standard errors imply that the estimate is likely to be close to the population proportion. Large standard errors imply that the estimate may be far from the population proportion.

In this section, we describe ways to sample households (that is, to select a sample of households). The survey will collect information (for example, labor status) from all children ages 5 through 17 in a sample of households. You sample households to identify children to interview for three reasons. First, lists of households are more readily available than lists of children (see the definition of "sampling frame," below). Second, once enumerators visit a household, the time required to collect data on another child in that household is usually small compared to the time needed to visit a new household. Third, the visit also enables you to collect household-level data.

Consider an example population of children ages 5 through 17 in Districts A, B, and C. Each district is divided into census enumeration areas (EAs; geographical areas used to help with counting in the census). Below, we describe ways to sample households from this population. Figure II.2, below, shows the population and an example of a sample.

Figure II.2. Households sampled from the population of interest



A **sampling design** (ILO Section 3.2.1) includes determining the sample size and the sample structure.

- The sample size is the number of households selected for the survey ($n_{households}$) and the number of children within those households (n)
- The sample structure is how those households are selected

The sample size and structure determine the standard error of estimates from the sample and the cost of conducting the survey. A larger sample size decreases standard errors and increases costs. Sampling households that are closer to each other increases standard errors and decreases costs, as we discuss below.

A sample is selected from the **sampling frame**, which is a list (or set of lists) that covers the entire population of interest. Below, we describe ways a sampling frame can cover the entire population of interest.

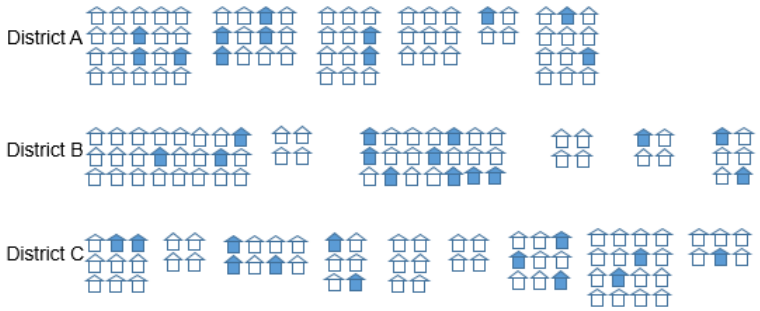
We begin by describing types of sampling designs and sampling frames. Next, we provide a basic type of sampling design and sampling frame. We mention an alternative design and provide references to the ILO manual for more information. Finally, we describe choice of clusters and strata (defined below) and of sample size.

B.1. Types of sampling designs and sampling frames

In **equal probability sampling**, all households have the same probability of being selected for the survey. This means that all children ages 5 through 17 have the same probability of being selected for the survey. With equal probability samples, you can estimate the proportion in child labor in the population just by calculating the proportion in child labor in the sample. (Non-equal probability samples require more complex calculations; they are generally preferred only when some parts of the population are of greater interest, causing oversampling in those parts.)

The simplest equal probability sampling design is **simple random sampling (SRS)** (ILO Section 3.4.1), which means sampling any $n_{households}$ with equal probability. Such a design could give, for example, the sample of households in Figure II.3.

Figure II.3. Households sampled by simple random sampling (SRS)



B.1.1. Non-SRS sampling methods

Instead of allowing *any* $n_{households}$ to be sampled (as in SRS), you might want to restrict samples in the following ways:

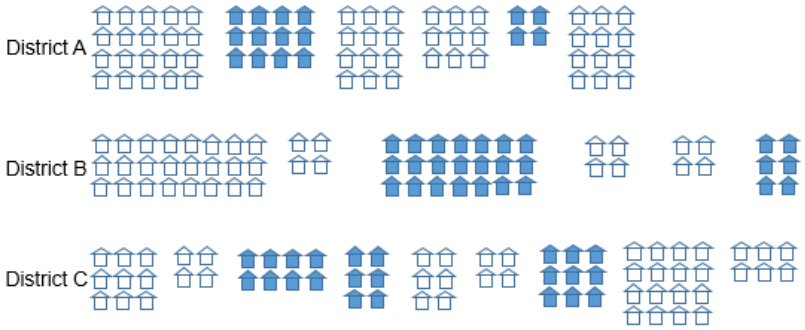
- Grouped into EAs (**cluster sampling**)
- Sampling more households from districts that have more households and fewer households from districts that have fewer households, so that the sample reflects the population in terms of districts (**stratified sampling**)
- Sampling households systematically from throughout the range of household sizes, so the sample reflects the population in terms of household size (**systematic sampling**)

Compared with SRS, these non-SRS methods allow for easier data collection (through sampling clusters of households that are near one another) and ensure that the sample resembles the population (through stratified or systematic sampling). A sampling design can combine any of these methods. Below, we describe these three non-SRS methods of sampling. In Section B.2, below, we describe how to combine these non-SRS methods into an overall equal probability sampling design. Researchers can refer to the ILO manual for information on more complex designs.

Non-SRS sampling method 1. Cluster sampling (ILO Section 3.4.3): Sampling groups (clusters) of households.

In any sampling design, **primary sampling units (PSUs)** are the first or only units that are sampled. In SRS, households are the PSUs, because they are the only units sampled. Households in the sample can be spread far apart from each other; by sampling groups (clusters) of households together as the PSUs, you can reduce travel time between sampled households. In the example population we introduced at the beginning of Section B, households are grouped into EAs. If you sample EAs instead of households, you get the sample shown in Figure II.4.

Figure II.4. Households sampled by one-stage cluster sampling



This is an example of **one-stage cluster sampling**, where there is only one stage (that is, one step) to the sampling ("Stage I").

Stage I: Sample clusters of households (PSUs)

Survey all households in the sampled clusters.

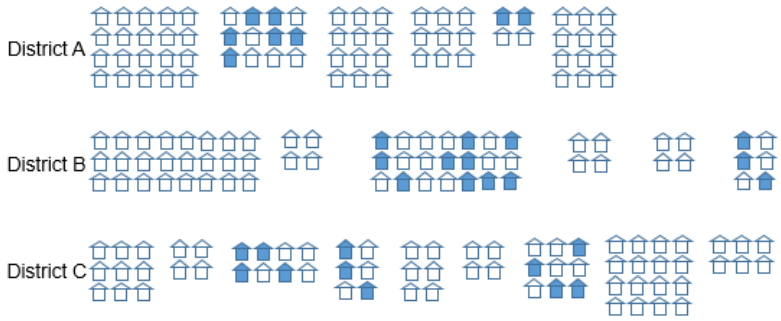
Another possibility is **two-stage cluster sampling**.

Stage I: Sample clusters of households (PSUs)

Stage II: In sampled PSUs, sample smaller clusters or individual households (secondary sampling units, SSUs).

Survey all of the sampled SSUs. Figure II.5 shows a sample selected through two-stage cluster sampling.

Figure II.5. Households sampled by two-stage cluster sampling



Similarly, three-stage cluster sampling is possible. For example, to reduce the travel time required among districts, you could first sample districts, then EAs, then households.

Cluster sampling has both pros and cons:

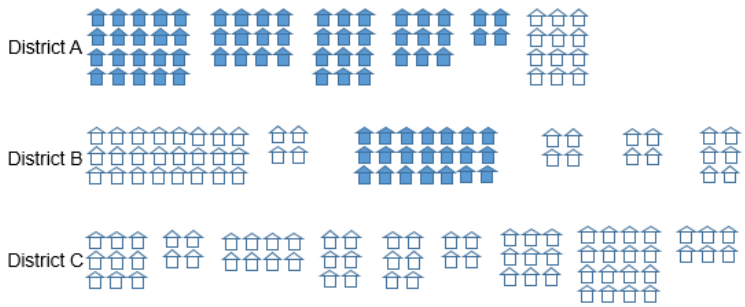
- **Pro**: It lowers costs in two ways: shorter travel between sampled households and less work to prepare a sampling frame (see Section B.1.2.).
- **Con**: Households in a cluster are similar, so you get redundant information. This increases standard errors relative to SRS of households.

In Section B.5, below, we provide a formula that balances these pros and cons for cluster sampling.

Non-SRS sampling method 2. Stratified sampling (ILO Section 3.4.2): Sampling separately within each group (stratum) of households.

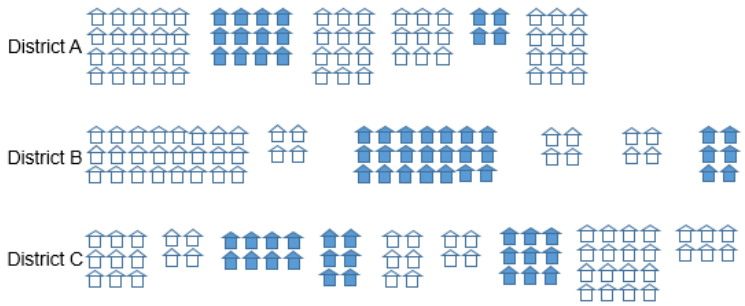
In our example, when randomly sampling EAs, by chance you could get a sample that does not include households from all districts (strata), as in Figure II.6.

Figure II.6. Households sampled by one-stage cluster sampling, not stratified on district



This sample does not look like the population because it does not have any households from District C. Instead, using stratified sampling, you could sample one-third of EAs in each district (stratum), as shown in Figure II.7.

Figure II.7. Households sampled by one-stage cluster sampling, stratified on district



Stratified sampling has some pros and cons:

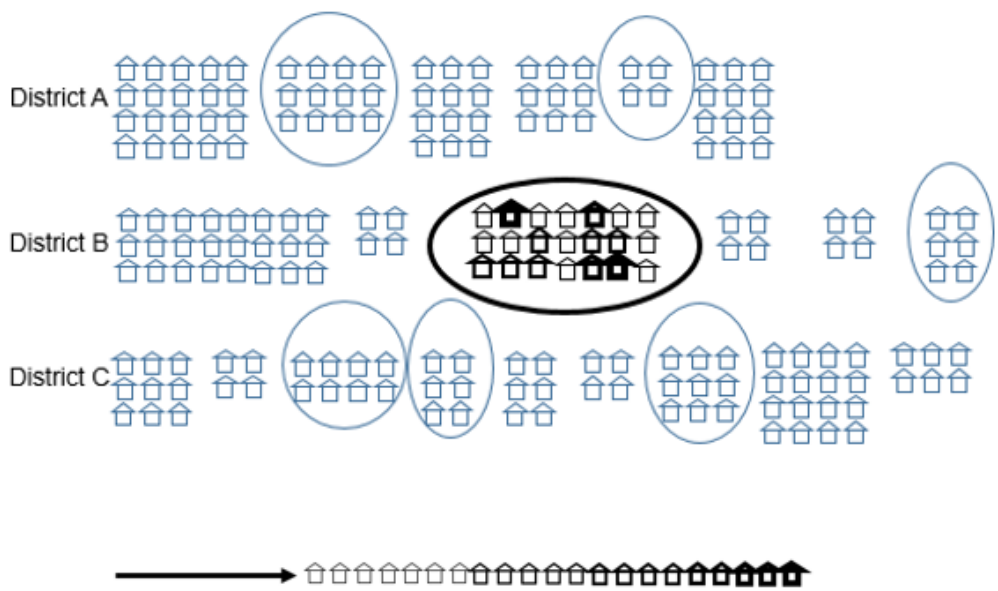
- **Pro:** It decreases standard errors by making the sample look more like the population in terms of strata variables.
- **Con:** It requires strata variables in the sampling frame (see below), more sampling steps, and more complex analysis.

Non-SRS sampling method 3. Systematic sampling (ILO Section 3.7.1): Sampling households equally spaced across a list, after a random start. This list may be ordered by some variables, which are called the "ordering variables."

Similar to stratified sampling, systematic sampling helps to make the sample more closely resemble the population. In stratified sampling, the sample looks like the population in terms of the strata variables (for example, districts). In systematic sampling, the sample looks like the population in terms of the ordering variables (for example, household sizes).

Suppose you want to sample $n_{households}$ out of $N_{households}$. First, list all $N_{households}$ in some order. For example, within a sampled PSU, you could order the households by their size, as in Figure II.8.

Figure II.8. Households ordered by household size, within a sampled PSU



Let $I = N_{households} / n_{households}$ be your sampling interval. Suppose I is an integer. For example, suppose you want to sample $n_{households} = 7$ out of $N_{households} = 21$ households. Then $I = 21/7 = 3$. Next, select a random integer r between 1 and I. Starting with r, calculate a sequence that keeps adding the interval I:

$$r, r+I, r+I*2 \ldots$$

Sample these households. In the example, suppose you randomly select $r = 2$. In that case you would sample households 2, 5, 8, 11, 14, 17, and 20 (every third household, starting with the second household).

Figure II.9. Households sampled by systematic sampling (after being ordered by size)

What if I is not an integer? For example, suppose you want to sample $n_{households} = 10$ households instead. Then $I = 21/10 = 2.1$. This time, select a random number (not necessarily an integer) between 0 and I as r. Calculate the above sequence, but round each number up to the nearest integer and sample these households. In the example, suppose you randomly select $r = 0.8$. Then the sequence is 0.8, 2.9, 5.0, 7.1, 9.2, 11.3, 13.4, 15.5, 17.6, and 19.7 and you would sample households 1, 3, 5, 8, 10, 12, 14, 16, 18, and 20.

Systematic sampling has pros and cons:

- **Pro:** It decreases standard errors by making the sample look more like the population in terms of ordering variables, as in Figure II.9. In our example, an SRS sample could by chance not include households of all sizes, as in Figure II.10. A systematic sample with size as the ordering variable prevents this.

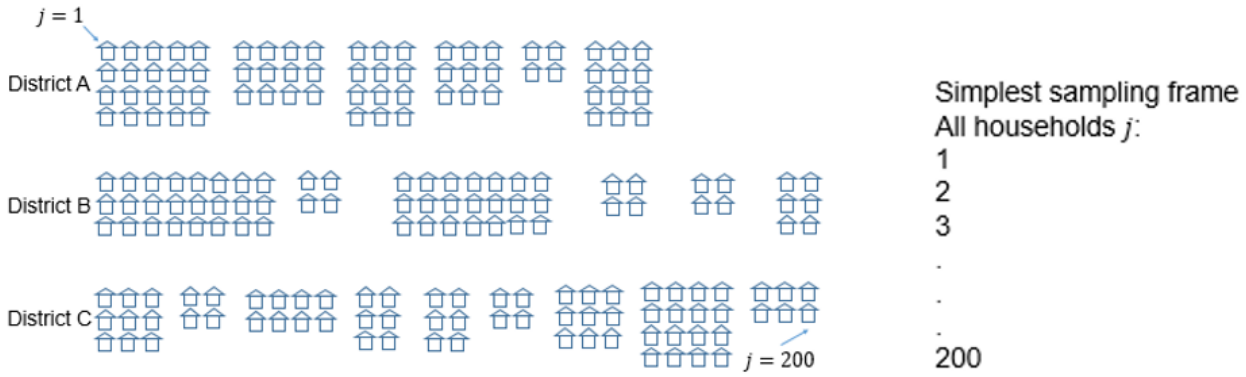
Figure II.10. Households sampled by simple random sampling (SRS)

- **Con:** Systemic sampling requires ordering variables in the sampling frame (see below).

B.1.2. Sampling frames (ILO Section 3.3.2)

The **simplest sampling frame** is a recent list of all households in the population. It is appropriate for selecting an SRS of households. You can obtain a list or create your own; however, creating a household list for large populations is costly.

Figure II.11. The simplest sampling frame

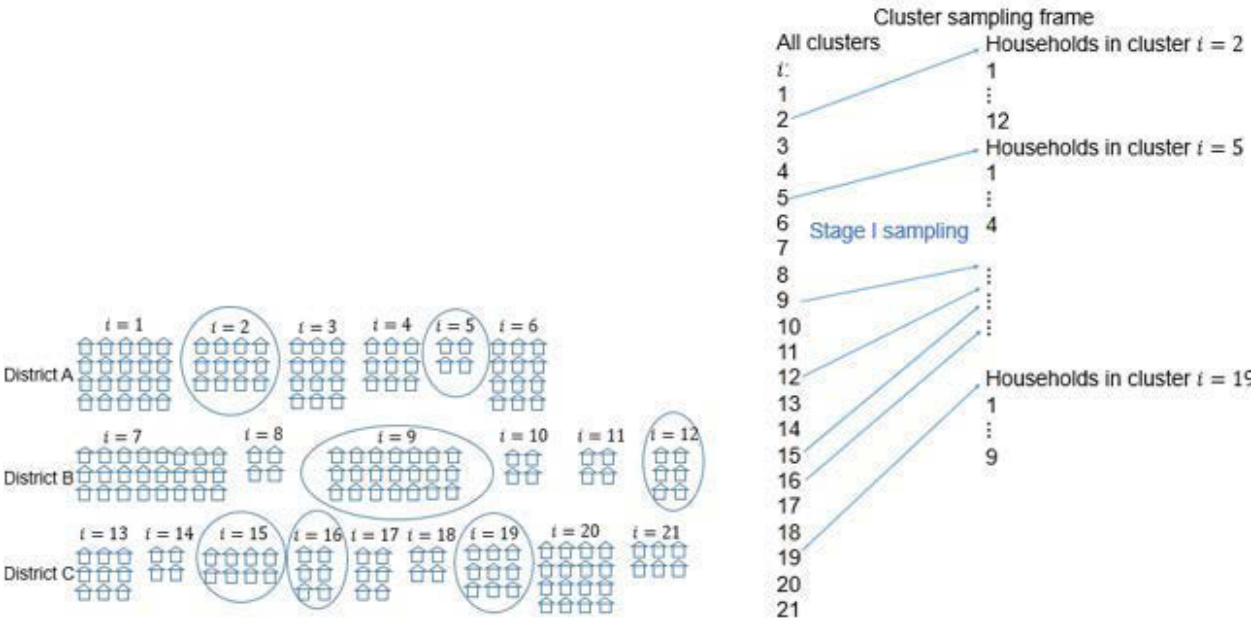


If an available household list covers the entire population of interest and is less than two years old, it can be used as a sampling frame. Otherwise, if the population of interest is small, you can consider creating your own household list. If the cost to do a household listing of the entire population is beyond the survey budget, a cluster sampling frame may be used.

A **cluster sampling frame** is appropriate for selecting cluster samples. It starts with a list of all PSUs (for example, EAs) in the population. For one-stage cluster sampling, this is enough. For two-stage or more cluster sampling, you need lists within the sampled clusters. The sampling frame must include a list for each stage of sampling.

For example, suppose you use two-stage sampling with EAs as PSUs and households as SSUs. In many cases, a list of EAs from a recent census is available from the country's statistical offices; this can be used for the first-stage sampling frame. Next, create household lists only within sampled EAs. The cost to list households is lower with the cluster sampling frame than with the simple sampling frame because you need to list households only within sampled EAs. Note: You could use old lists of EAs, but household lists older than two years may be out of date and inaccurate.

Figure II.12. A cluster sampling frame



The processes of developing the sampling design (how households are selected) and sampling frame (the lists from which households are selected) are linked. Choice of clusters or strata can be based on what information is in available sampling frames. Sampling frames can be developed from multiple sources. For example, one source may list clusters by district whereas another source may list clusters by urban versus rural. These sources can be combined and used to stratify by district and urban versus rural.

You cannot learn about the population if there are serious problems with the sampling frame. Three types of potential problems with sampling frames are duplications, undercoverage, and households that are outside the population:

- **Duplications:** A household or cluster is listed more than once.
- **Undercoverage:** Households in the population are not represented in the sampling frame. For example, if the population of interest is current households, a sampling frame listing households from five years ago would miss new households.
- **Elements outside the population:** Households outside the population are represented in the sampling frame. For example, an old sampling frame may include households that have since moved away. Enumerators would be unable to find these households at data collection time, reducing sample sizes and increasing effort required for data collection.

You can avoid problems with the sampling frame by using credible sources of lists or spending sufficient resources to create new household lists. Census offices or national statistics offices are

usually credible. Local leaders can help create lists or update existing lists. Lists more than two years old will likely require updating. Remove duplications in the sampling frame before sampling.

B.2. A basic type of sampling design and sampling frame (ILO Sections 3.2.5 and 3.3.2)

Here we describe stratified two-stage cluster sampling that is designed to be equal probability and to minimize costs.

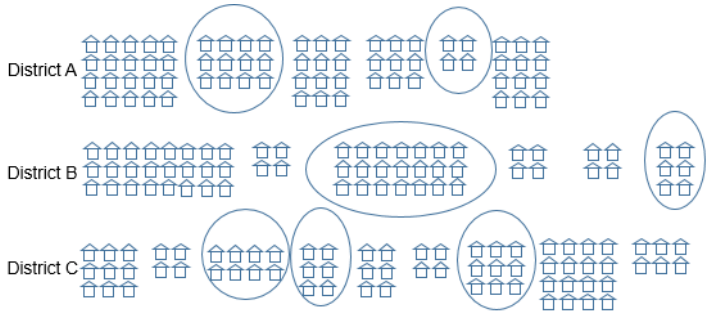
Stage I: Sample PSUs with equal probability

Stage II: In each sampled PSU, sample a fixed fraction (for example, 50 percent) of households with equal probability

At both stages, the equal probability sampling can be simple (SRS), stratified, or systematic.

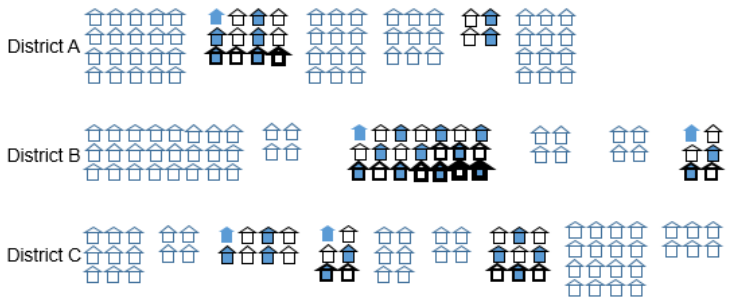
For example, at Stage I, you can stratify on district to help make the sample similar to the population in terms of districts. To make the sampling equal probability, take the same fraction of PSUs in each district (for example, one-third); see Figure II.13.

Figure II.13. Stage I of the basic sampling design



Then, at Stage II, you can order households within PSUs by their size and do systematic sampling to get a good spread across sizes. To make the overall sampling equal probability, take the same fraction of households in each sampled PSU (for example, one-half); see Figure II.14.

Figure II.14. Stage II of the basic sampling design



This design uses a cluster sampling frame. At Stage I, researchers can use existing lists of PSUs (that may be several years old), and then create household listings only within sampled PSUs (or obtain them, if new ones are available).

The basic sampling design works for many child labor research purposes. However, in some cases there are reasons not to use it:

- A recent household listing of the entire population of interest is available and there are sufficient resources to spend on travel among sampled households across the population. In this case, one can use a nonclustered sampling method (households as PSUs).
- A recent measure of PSU sizes is available and the researchers are very familiar with probability proportional to size (PPS) sampling of PSUs or are working closely with a statistician. See Section 2.3, below.

B.3. An alternative type of sampling design and sampling frame (ILO Section 3.6)

A popular alternative equal probability design is the following:

Stage I: Sample PSUs with PPS[1]

Stage II: In each sampled PSU, sample a fixed number (for example, 10) of households with equal probability

At both stages, the sampling can be stratified or systematic (and at Stage II, the sampling can be SRS). Although sampling of PSUs at Stage I is not equal probability, after both Stages I and II, the sampling of households is equal probability.

This design also uses a cluster sampling frame, but with an important difference from the basic equal probability sampling design. To do PPS sampling at Stage I, the list of PSUs has to include a recent

[1] In Section B.2, we describe sampling the same fraction of PSUs in each district (for example, one-third). In other words, the number of PSUs sampled is proportional to the number of PSUs in each district. This is not PPS sampling. In PPS sampling, each PSU is sampled with probability proportional to the number of households in the PSU. In other words, a bigger PSU with more households is more likely to be included in the sampling frame than a small PSU with fewer households.

measure of the sizes of the PSUs (the number of households in each PSU). This may be difficult for most researchers to obtain. For more information on PPS sampling, see the ILO manual or consult a statistician.

B.4. Choice of PSUs and strata (ILO Sections 3.4.2 and 3.4.3)

PSUs must have clear boundaries. Census EAs are a good example. Villages or communities often have less well-defined borders. Using larger PSUs (for example, districts or regions) will increase standard errors more than using smaller PSUs (for example, EAs). Because fewer PSUs are selected than households, it is more important to stratify sampling of PSUs to get a good spread of them across strata variables. Possible strata variables include regions, districts, or urban versus rural. For example, strata can be defined by both district and urban versus rural: District A–urban, District A–rural, District B–urban, District B–rural, District C–urban, and District C–rural.

B.5. Sample size and power calculations (ILO Section 3.5)

Choice of sample size is a balance between reducing standard errors and taking into consideration the timing of the survey and the budget. This choice is usually focused on one key quantity of interest, for example, the proportion in child labor. To calculate sample sizes, consider the desired standard error (*se*), the predicted variability (*v*), and the design factor:

The **design factor (deft)**[2] is the ratio of the standard error from the sampling design to the standard error that you would get from an SRS of children[3] (ILO Sections 3.5.2 and 7.8). For example, $deft = 2$ indicates that the standard error from the sampling design is twice the standard error from an SRS of children. Systematic sampling or stratified sampling to improve spread across key variables can reduce standard errors, giving a deft of less than one. However, to make sure you get large enough sample sizes, consider only increases to the design factor. For equal probability sampling, the deft increases mostly from cluster sampling at Stage I (sampling PSUs).

Calculate the design factor from PSU size and the similarity within PSUs, measured by the intra-cluster correlation (ICC). If *b* is the average number of households sampled per PSU and AveHH is the average number of children ages 5 through 17 per household, then (ILO Sections 3.5.3 and 7.8):

$$deft^2 = 1 + (b * AveHH - 1) * ICC$$

For noncluster sampling (that is, households are the PSUs), $b = 1$ and the ICC measures similarity of children within households.

---

[2] Readers might be familiar with the **design effect (deff)**, which is equal to the square of the design factor (deft): $deff = deft^2$.

[3] This text does not consider an SRS of children. Instead, we have assumed that households are sampled, and all children in sampled households are surveyed.

The total number of children to be surveyed is:

$$n = \frac{v}{se^2} * deft^2$$

Dividing n by $AveHH$ gives $n_{households}$, the total number of households to be sampled.

Choosing the intra-cluster correlation, ICC (ILO Section 7.9)**:** Use ICC = 0.2, the highest ICC mentioned in the ILO manual. A high ICC assumes children in a PSU are similar, giving redundant information. This gives higher standard errors than an SRS of children, increasing the deft. Using a high deft increases the sample size requirements.

For cluster sampling, choosing the average number of households sampled per PSU, b (ILO Section 7.9)**:** Researchers consider the cost of visiting a new PSU (C_1) compared to the cost of surveying another household (C_2) in an already visited PSU, and use the formula[4]:

$$b = \sqrt{\frac{C_1}{C_2}\frac{1-ICC}{ICC}}$$

C_1 includes the cost of travel to a new PSU and the cost of obtaining a list of households within the new PSU if the sampling design is a two-stage design.

Choosing the predicted variability, v (ILO Section 3.5.2)**:** Most quantities of interest are averages (for example, the average number of hours worked per week) or proportions (for example, the proportion in child labor). For an average, choose variability to be:

$$v = \left(\frac{\max - \min}{4}\right)^2$$

Where max is the maximum response you expect to see (the highest number of hours worked per week) and min is the lowest response you expect to see. For proportions, if p is the proportion you expect to see, take the variability to be:

$$v = p(1-p)$$

Choosing the desired standard error, ***se***. The standard error measures how close you expect the estimate from the sample to be to the true population value. To choose the desired standard error for

[4] We use ICC = 0.2 both to describe similarity of children within PSUs and similarity of households within PSUs.

a key quantity of interest, consider surveys done in similar contexts. How large is the typical standard error for these surveys? What level of precision is the audience of the reported results going to expect?

Sample size calculations often start with a choice of desired standard error and arrive at a necessary sample size, which in turn implies a survey budget. However, projects often have a maximum budget before sampling plans are developed. Therefore, a researcher can specify a desired standard error, calculate the implied budget, and assess whether that standard error is too ambitiously low and needs to be increased.

C. Step 3: Collect data

We do not fully describe data collection here (see Chapters III and IV on survey development and data formatting and validity for more detail on collecting and working with data). In this section, we focus only on **nonresponse**, the failure to get survey responses from sampled households or people (ILO Sections 3.2.3 and 7.1.2). Reasons for nonresponse include the following:

- An enumerator was unable to locate a sampled household or person
- A sampled household or person did not consent to be surveyed
- A sampled household or person did not answer certain questions

Nonresponse reduces sample sizes, increasing standard errors. More important, people or households that do not respond may be different from those that do. For example, households that do not consent to be surveyed may be poorer than those that do, or may have more children in child labor and be more reluctant to participate in the survey. This makes the sample look different from the population. Strategies to reduce household nonresponse include scheduling a survey time with sampled households or visiting them multiple times.

Additionally, people who do not answer certain questions may be different from those who do answer those questions. Pre-testing questions helps to ensure that interviewers are trained to make respondents comfortable answering all questions.

D. Step 4: Analyze and report the data

We describe the steps to analyze your data and to report on it in Chapter V on analysis and reporting.

Exhibit II.1. Sampling Worksheet

This worksheet includes questions to help researchers work through defining their population of interest (Step 1) and developing a sampling design and sampling frame (Step 2). Read the text describing Steps 1 and 2 before completing this worksheet.

We provide example answers in italics.

Worksheet Part A: Basic Type of Sampling Design

STEP 1: Define the population of interest

1.1 What is the geographic target area where you are implementing your project and thus that you want to study?

Example Case: *Districts A, B, and C*

Your answer: _____

1.2 What is the time period you want to study in this survey?

Example Case: *We need to know the baseline outcomes from before we implement the project in April 2018. Therefore, we will conduct the baseline survey from January to March 2018.*

Your answer: _____

_____

1.3 Following the International Labour Organization age range for defining child labor, the population of interest includes children ages 5 through 17. Are there any additional groups of people that you want to study?

Example Case 1: *No*

Example Case 2: *Yes, in addition to all children ages 5 through 17, our project is interested in outcomes for youth ages 18 through 22.*

Your answer: _____

_____

STEP 2: Develop a sampling design and sampling frame

2.1 Will you use the basic type of sampling design and sampling frame, which we described in the text in Section B.2 "A basic type of sampling design and sampling frame"? Check one:

☐ Yes → go to 2.2
☐ No → go to 2.8

Stage I: Sample primary sampling units (PSUs) with equal probability

2.2 What are your PSUs?

Example Case: *Census enumeration areas*

Your answer: _____

2.3 Stage I sampling frame: How will you obtain a list of PSUs?

Example Case: *We will obtain a list of census enumeration areas from the country's statistical office.*

Your answer: _____

_____

Your PSU sampling frame might include information about PSUs that you can use as strata variables for stratification sampling or as ordering variables for systematic sampling. Strata variables have values that are categories (for example, districts); ordering variables have values that are numbers (for example, PSU sizes). Below, we give some options.

2.4 Will you use stratified sampling to sample PSUs?

☐ Yes → Check any variables below that you want to use for stratifying PSUs and that are included in your PSU list. Checking more than one box means that strata use both variables. For example, checking both "PSU district" and "PSU classification into urban/rural" gives these strata: District A–urban, District A–rural, District B–urban, District B–rural, District C–urban, and District C–rural.

- ☐ PSU region
- ☐ PSU district
- ☐ PSU classification into urban/rural
- ☐ Other(s):

☐ No → go to 2.5

_____

2.5 Will you do systematic sampling of PSUs? Check any variables that you want to use for ordering in systematic sampling of PSUs and that are included in your PSU list:

- ☐ PSU size
- ☐ Other(s):

_____

Stage II: In each sampled PSU, sample a fixed fraction of households with equal probability

2.6 Stage II sampling frame: after Stage I sampling, how will you get household lists in sampled PSUs?

- ☐ From an existing source.
- ☐ [If you check this box, what is your source for household lists? When was it created?]

Example Case: *We have lists of households in each enumeration area from the 2017 census.*

Your answer: _____

_____

- ☐ [Year created (cannot be more than two years old): _____]
- ☐ We will create household lists.
- ☐ [If you check this box, what is your procedure to create household lists?]

Example Case 1: *We will send enumerators to visit all households in sampled PSUs. At each household, they will determine whether there are children between the ages of 5 and 17 in the household after explaining the purpose of their visit.*

Example Case 2: *We will send enumerators to speak with village leaders to obtain lists of households with children between the ages of 5 and 17 in the sampled PSUs.*

Your answer: _____

_____

_____

_____

_____

Your household lists may include information about households that you can use as ordering variables in systematic sampling. For simplicity, we do not describe stratified sampling of households. For more

information on stratified sampling, we direct readers to the ILO manual Sampling for Household-Based Surveys of Child Labour.

2.7 Will you do systematic sampling of households? Check any variables that you want to use for ordering in systematic sampling of households and that are in your household list:

- ☐ Household size
- ☐ Other: _____

[If you answered "yes" to question 2.1 and completed questions 2.2 through 2.7, skip question 2.8 and go to the calculations of sample size.]

Worksheet Part B: Alternative sampling designs

2.8 Why are you choosing an alternative type of sampling design and sampling frame?

Example Case: *1: We have a household listing from 2017 that covers all households in the geographic target area. We have sufficient resources to travel among households across this whole area, so there is no need to do cluster sampling.*

Example Case: *2: We have data on census enumeration area sizes from 2017, so we would like to use this to sample enumeration areas with probability proportional to their size (PPS sampling) at Stage I.*

Your answer: _____

_____

_____

_____

_____

2.9 Describe your sampling design.

Example Case: *1: We will take a simple random sample (SRS) of households from the population.*

Example Case: *2: We will do PPS sampling of enumeration areas at Stage I, followed by sampling 10 households per enumeration area.*

Your answer: _____

_____

_____

_____

_____

2.10 Do you have a statistician or statistical consultant to help develop this alternative sampling design?

☐ Yes
☐ No → consider using the basic design

Worksheet Part C: Sample size calculations for the basic type of sampling design

2.11 Will you use the basic type of sampling design and sampling frame, which we described in the text in Section B.2 "A basic type of sampling design and sampling frame"? Check one:

- ☐ Yes → complete sample_size_template.xlsx to perform sample size calculations.
- ☐ No → consult the ILO manual, other resources, and statisticians to perform sample size calculations.

III. SURVEY DESIGN AND DATA COLLECTION

This chapter covers survey design, including drawing on existing surveys, translating survey instruments, using skip patterns correctly, defining households and survey respondents, capturing age data, targeting survey questions to children, and building in quality checks.

A. Draw from existing surveys

When developing surveys, draw from existing, internationally recognized surveys, making appropriate adaptations for the local context. Existing surveys can provide a framework to ensure the inclusion of relevant and appropriate questions. Using survey questions that researchers have already successfully fielded reduces survey development and pre-testing effort. Drawing on widely used surveys can also allow for the comparison of your data with data from other surveys. This allows researchers to learn about how child labor is changing over time or to compare child labor statistics from project areas to national statistics or to other countries.

When drawing on existing surveys, make sure that they are relevant to the local population. Before using questions from existing surveys, determine if it makes sense to adapt the questions to the local context. If the questions relate to concepts that differ across cultures, it might be necessary to adapt them so the local population will understand them. If you plan to compare new data to previous findings, do not adapt the questions (provided they are understandable) so as to ensure comparability between data sources. Pre-test your surveys—whether adapted or not—in the field with populations similar to the target population. Pre-testing involves testing the survey either by surveying respondents or reviewing items in a focus group. Piloting a survey typically includes practicing the survey as well as other data collection procedures, such as finding households and describing the survey. Pretesting ensure that questions are appropriate for the local context and will elicit the desired information. Survey documentation should make note of any surveys used as a reference, as well as any adaptations.

The following surveys are available online. Adapt survey questions when necessary and cite the sources of questions used:

- For questions on child labor:
  - The national statistics office in your country may have conducted a National Child Labor Survey (NCLS). These are often done as part of a national living standards survey. Questionnaires, reports, and data sets are available on ILO's website. You may wish to contact the national statistics office in your country for further support and information on the NCLS in your country.
  - ILO has also produced model child labor survey questionnaires, which are stand-alone child labor surveys that are adaptable for global use (known as the SIMPOC surveys).
- For questions on socioeconomic status, demographics, and health:
  - The World Bank's Living Standards Measurement Surveys (LSMS) are used around the world to measure living standards. These surveys and data from many of them are available online.
  - USAID's Demographic and Health Surveys (DHS) are nationally representative household surveys that focus on population, health, and nutrition. UNICEF's Multiple Indicator Cluster Surveys (MICS) are conducted around the world, collecting data on key indicators on the well-being of children and women.

B. Use translation procedures that maintain original meaning across languages

When translating existing survey questions into local languages, translate and pre-test carefully so the meaning of questions is not lost. Keep question formats similar and maintain the same measurement properties, including the range of response options, particularly if your research team wishes to compare the data you collect to data from other surveys. Develop a translation protocol (a written plan), follow the protocol, and describe it in survey documentation.

Translate all elements of a survey that will be read to the respondent into the respondent's language before administering the survey. Thorough translation reduces the need for spontaneous interpretation and ensures that survey questions are consistent for all respondents. Effective translation captures the conceptual equivalent of each question without necessarily translating the original text word-for-word. When translating, stick to simple language targeted at the relevant audience and avoid technical or slang terms some might not understand.

Take steps to ensure that questions are consistent for all respondents even if complete translation is not feasible. In some cases, spontaneous interpretation will be unavoidable—if, for example, the number of respondents who require translation into a language is too small to justify the cost of translating the full survey, or if the language is not a written language. In such cases, include guidance for enumerators in the translation protocol. For nonwritten languages, enumerators can listen to recordings of translations during training so that questions are standardized. Alternatively,

enumerators can develop a vocabulary list of key terms in the target language so that all enumerators are using the same translations for those terms, especially terms essential for estimating the prevalence of child labor. Be sure that the vocabulary list includes words used to describe work schedules and working conditions (you may wish to include others). If an interpreter is used, state in the guidance who should interpret for the respondent and what steps the enumerator should take to ensure that the interpreter understands and accurately conveys the meaning of each question. Train enumeration teams to consider confidentiality concerns in cases where a community member known to the respondent household offers to interpret and the respondent must therefore share his or her responses with that individual.

Use a team translation approach to translate and pre-test surveys. In this approach, two translators produce independent initial translations, then the research team reviews and discusses the translations to agree on a version to send to the field for pre-testing. Pre-testing translations includes discussing each question with respondents, such as asking them to repeat questions in their own words and asking them if there were questions they did not understand. If an in-depth pre-test of the entire survey is not feasible, focus your testing on sensitive questions, questions that were difficult to translate, and ones that are key to measuring child labor and other required outcomes. Translation can be an iterative process involving modifications during review and pre-testing. Document all steps of the translation process.

C. Use correct skip patterns

Use skip logic to control the flow of a survey and direct the correct questions to survey respondents, based on responses to previous questions. For example, when using skips, if a respondent answers "no" to "Are you currently enrolled in school?" use a prompt to direct the enumerator to skip the subsequent question, "What grade are you currently enrolled in?" Clear and logical skip patterns ensure that respondents answer all relevant questions so data are complete while ensuring they do not answer irrelevant ones, wasting time and creating confusing and contradictory data. When respondents answer questions that should have been skipped, the data are difficult to interpret; for example, if the data report in question 1 that a respondent does not attend school and in question 2 that he or she is in 8th grade, data users will not know which answer is correct.

When designing a survey, identify where to place skip patterns to ensure logical question flow, and provide clear instructions for enumerators to follow. It is easiest to follow skip patterns if they prominently mark which responses lead to which questions. Checking the survey thoroughly is vital to make sure all skip patterns are correct. A good way to test surveys for correct skip patterns is to create a mock respondent, answer the survey as that respondent, and check that all relevant questions were answered, with no answers for irrelevant questions. Be sure enumerators are well trained on skip logic. Pre-test the surveys in the field and check skip logic in the data from pre-testing to ensure that enumerators are correctly following skip patterns.

Use electronic data collection whenever possible to reduce data collection errors. Electronic data collection programs automate programmed skip patterns, eliminating errors that arise from enumerator error. Designing and programming the electronic survey instrument requires careful attention and thorough testing to make sure all skip patterns are correct. An error in skip logic programmed into the electronic survey will be made every time the survey is conducted. That is, whereas enumerators may occasionally follow skip logic incorrectly, if skip logic is programmed incorrectly in an electronic survey, the program will follow the incorrect logic every time the survey is conducted.

D. Clearly define "household" and "head of household"

Use clear and precise definitions of a household and a household head to ensure that survey designers, enumerators, respondents, and analysts have a common understanding. A clear definition of household also ensures the consistent identification of children who are considered part of the household, including foster children, children working in the household, and children who might live in but not be present at the home on the date of the survey. Include a clear definition of household in the survey instrument and enumerator training documents and instruct the enumerator to read the definition to respondents before asking them to identify household members.

Because the definition of a household varies depending on context, consider using other surveys, such as the national census or DHS, to identify the most commonly accepted and complete definition of household within the country. The standard definition of a household used around the world by the World Bank's LSMS is a group of people who live together, pool their money, and eat at least one meal together each day (United Nations 1989). People who live together but do not pool resources—for example, live-in domestic helpers or people who live together to reduce housing costs—typically are not considered members of the same household. A complete household definition specifies the length of time an individual must have been present to be considered a member (for example, 3 of the past 12 months), as well as any exceptions to this rule, such as a newly married spouse or a newborn baby. In some contexts, the definition of a household may not be obvious, for example, polygamous families in which wives have separate dwellings. In such cases, choose an appropriate definition and apply it consistently.

E. Clearly identify targeted respondents

Specify which respondent(s) should answer each question or set of questions and include text in the survey instrument explaining who should respond to each set of questions. Surveys for identifying child labor often include modules targeted at the household head, children's caretakers, and children. For modules targeted at an adult but about children, the children's primary caregiver is typically a better source than the household head because she or he is more familiar with details about the children. Devise a strategy to gather data on all children ages 5 through 17; this may include survey questions for all children in those age ranges. Clearly specify the target age range

for modules about or directed at children and in each module specify whether the child should respond or if an adult can answer on the child's behalf. See discussion below on writing age-appropriate questions for children.

F. Give instructions for dealing with absent respondents

Provide guidance on how to proceed if a person who is supposed to be interviewed is absent. In survey instructions, specify how many attempts enumerators should make to interview the absent household member and instruct enumerators to collect contact information and/or request a better time to return to interview that person. Instructions may vary depending on who is absent. For example, if the household head is absent, instruct enumerators to return when the individual will be home. If one of the children is absent, instruct enumerators to complete the other modules of the survey and return later to interview that child. If a person cannot be interviewed, provide clear guidance in the survey instructions for what the enumerator should do—possibly interview the most knowledgeable other household member in place of an absent household head or absent child (a proxy respondent) and note the absent household member. In the survey, include a place to note when a proxy respondent completes the interview in place of a targeted household member. If it is not possible to interview the full household—for example, if the entire household cannot be located, or if all members refuse to participate—instruct enumerators to make notes about why the household did not participate. Capturing and analyzing data on why households did not participate can be helpful in planning for future surveys.

G. Capture accurate age data

Accurate information about age is fundamental to identifying child labor because the definition of child labor is based on the age of the child. In the enumerator instructions, include strategies for estimating age data if respondents do not know their own exact age or the exact ages of others in their household. Some strategies include asking for the date of birth, asking to see identification or birth documents, or triangulating information based on other family members' ages and/or major events. To anchor a birth year around a major event, the enumerator can ask questions such as who was president when a person was born, or ask if the person was born when a major event, such as an earthquake or hurricane, occurred. This approach requires enumerators to place such events in time accurately; cover this method in detail during enumerator training if it will be used. Similarly, ages can be triangulated around other family members' ages by asking how old parents or older siblings were when a child was born, started school, and so on. Include a field in the survey for the enumerator to indicate how he or she collected or estimated age data.

H. Consider appropriateness of questions for child respondents

Interviewing children requires special care. Young children may be unwilling to talk to strangers, absent-minded during an interview, unable to comprehend questions, or unable to recall past events accurately. Careful survey design and enumerator training are necessary to protect children and to collect accurate data from them. Hire enumerators who are experienced in interviewing children and

who have passed a recent background check. In enumerator training, address techniques for interviewing children and particularly how to ensure children's safety and comfort.

Ensure child safety and well-being during and after the interview. It is crucial for children's protection as well as for the accuracy of the data that child respondents be safe and feel safe giving accurate responses to questions without fear of repercussions. Instruct enumerators to conduct interviews out of earshot of parents or others who could influence or negatively react to children's responses. Parents might be wary of leaving their child with a stranger, so instruct enumerators to look for a location where children can be seen but not heard while the interview takes place. Keep children's responses confidential; the only exception is when reporting suspected child abuse or endangerment. Know requirements for reporting such cases in your country and develop protocols for reporting to authorities. Parents must provide permission for their children to participate in the survey and children must assent (provide affirmative agreement) to participating in the survey. Children must understand what they are agreeing to.

Keep survey questions for children simple and use concepts that children understand. These are key strategies for developing questions for children:

- Keep questions short.
- Give few response options. As a rule of thumb, limit response options to three for children younger than age 10 (Holaday and Turner-Henson 1989).
- Avoid using concepts or technical terminology—such as "business partner," "economic activity," "piece rate," "overtime pay," and "accident insurance"—that are difficult for children to understand. If you are asking children about these concepts, describe them in simple language. For example, to learn whether children are paid through a piece rate, you could ask them about how they are paid and if they know whether the amount they are paid depends on how much they produce or the time they work.
- Help children estimate hours worked. Provide visual aids, using sunrise, meal times (if standard), and sunset as familiar markers to help children describe when they start and stop activities so that you can calculate the time they spend on different activities.
- Help children estimate weight when asking about carrying heavy weight. Young children are unlikely to know what things weigh, but they might be able to compare the weight of things they lift to other things in their environment. Train enumerators to convert these values to standard weight measurements such as kilos.
- Pay special attention to questions about sensitive topics, such as physical or sexual abuse, illegal work activities, and so on. Tailor sensitive questions to different age groups using language that children will understand and with which they are comfortable.

Table III.1 provides examples of questions that may be challenging for children and alternatives that may be easier for children to understand and answer.

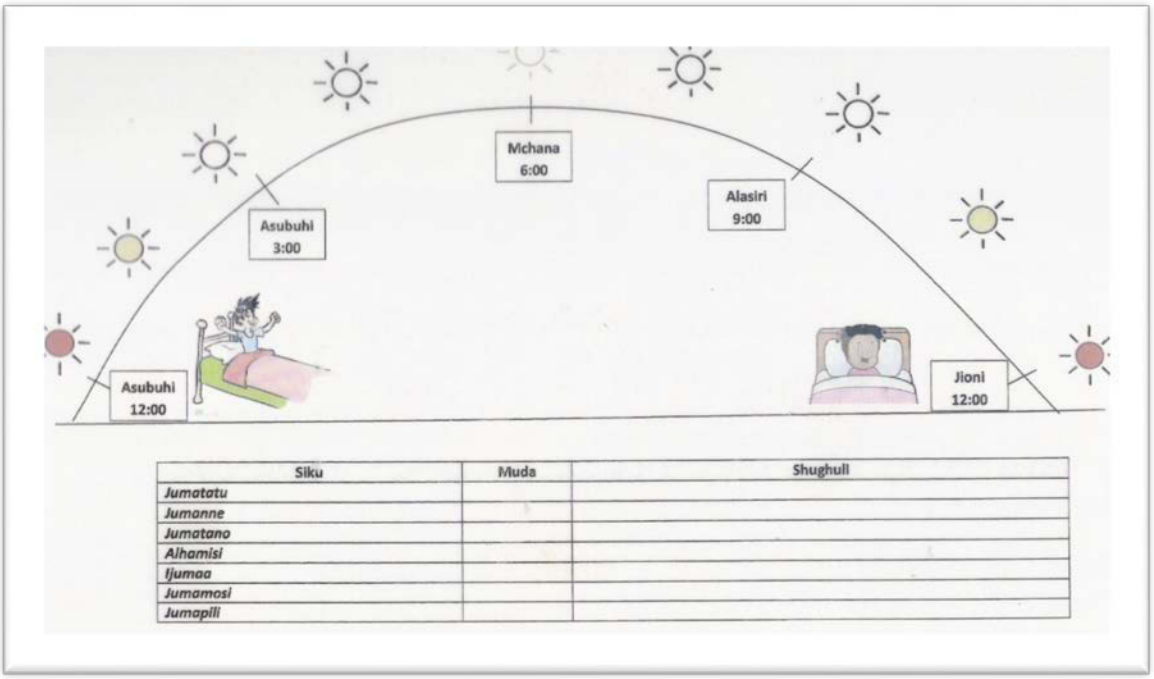
Table III.1. Challenging and easier questions for young child respondents

Challenging questions for young child respondents	Easier questions for young child respondents	Notes
1. What was the mode of payment for the last payment period? a. Piece rate b. Hourly c. Daily d. Weekly e. Monthly f. Upon completion of task b. g. Other (specify)	1. **When they pay you, how do they figure out how much to pay you?** a. By how much you work or make? [Provide an example appropriate for their type of work, like "for how many candy you sell?"] SKIP to 3 b. By how long you work? c. Other (specify) d. Don't know 2. **Do they pay you based on how many hours, days, weeks, or month you work?** a. Hours b. Days c. Weeks d. Months e. Other (specify) f. Don't know	Young children are likely to struggle with the response options in the original question, like "piece rate" or "hourly," but could tell you if their pay is based on quantity produced or time worked. Use simple terms children will know. The optimal number of response options for young children is two to three, although more options may be required in some cases. The original question has many response options and will be easier to answer if broken into two questions. It is important to include "don't know" options because if respondents do not know how they are paid—as is likely to be the case with young children—they will be unable to answer this question, no matter how simply it is phrased.

Source: Original survey questions are drawn from the annotated survey at the end of this chapter. Revised questions were written by the authors.

A key component of child labor surveys is to learn about how many hours children work. Older children will be able to list their work hours precisely in many cases. However, children under ten, will not in some cases. A visual aid, such as the one shown in Figure III.1 can help young children describe their work schedule in relation to key events, such as sunrise, sunset, and meal times. Although this may seem imprecise, for young children it may be more precise than asking them to give specific times.

Figure III.1. Visual aid for time-related questions for young children



Source: OCFT grantee visual to accompany child survey.

For more guidance, de Leeuw et al. (2004) provide details about issues that arise when developing surveys for children.

Consider using different versions of specific questions for children of different ages. Older and younger children are likely to be involved in different types of activities, and children of different ages will have different levels of comprehension. Carefully target different age groups with questions using appropriate vocabulary, concepts, and assistance when necessary. Grantees could also provide enumerators with visual aids or cards to help young respondents understand more difficult questions.

Pilot testing is crucial—especially when surveying children. Pilot questions for children with children of the same age range. The resource by de Leeuw et al. (2004) also outlines strategies for piloting surveys with children. Methods include using focus groups, cognitive interviews (a sort of pre-test that involves asking respondents what they think the survey questions are asking), or observing survey administration. When observing surveys in a pilot, note what the enumerator and respondent say, to look for clues that questions might benefit from revision. (Does the enumerator have to stray from the questions often to help the respondent understand? Does the respondent interrupt to ask for clarification? Does the respondent appear confused or distressed?)

For survey questions targeting young children who might have problems comprehending or answering questions, specify whether an older household member may assist. Parental involvement can influence responses, so be sure guidelines for parental involvement are consistent

for all respondents. Include in the survey a place where the enumerator can record whether other family members were present for an interview and whether the child received help responding.

I. Use data collection procedures that ensure high quality data

Select vetted, experienced enumerators; train them thoroughly on the specifics of the survey instrument; and test their ability to administer the survey correctly. Competent, well-trained enumerators are key to collecting high quality data. Hire enumerators with experience interviewing populations similar to those targeted by your survey and knowledgeable about the topics covered. Whenever possible, have enumerators complete a background check to be sure they do not have a criminal background. In enumerator training, include general interviewing methodology—including how to conduct interviews with children—as well as survey-specific training to understand the survey instrument and specific survey questions. Test enumerators to ensure that they understand the survey well before fielding the survey. In some contexts, it would not be considered appropriate for enumerators to survey respondents of a different gender; when hiring and assigning enumerators, consider whether this is a factor in your context. If the gender of the enumerator is important, you may need to send enumeratos in mixed-gender pairs. For detailed guidance on enumerator selection and training, the University of Michigan's Cross-Cultural Survey Guidelines has a detailed chapter on the topic (also linked in the online resources list).

Monitor data collection as the survey is in progress. While surveys are being conducted, communicate with the data collection team to monitor and report on response rates, including the reasons for nonresponse. Document information about how an interview was performed, including whether enumerators used interpreters and/or proxy respondents, whether they estimated age information, and other details. Instruct enumerators to review every survey immediately after completing it to make sure it is complete, and to check for errors in recording responses. Instruct supervisors to review surveys for completeness and logical consistency and follow up with enumerators as soon as possible to attempt to address errors. Also instruct supervisors to call or visit some households (typically 5 to 10 percent) to verify responses to be sure that enumerators are not inventing data. For enumerators whose data are suspicious, 100 percent of cases should be verified by supervisors. Be sure enumerators inform respondents that they may receive a call or a visit from a supervisor to talk to them about the survey.

Run diagnostic tests on the data in real time to look for errors. Review data to look for incomplete surveys, implausible or impossible values, or other inconsistencies. If using computer-assisted personal interviewing (CAPI), data can be reviewed in real time. If using paper surveys, concurrent data entry allows for data reviews while data collection is still in progress.

See Chapter IV on data format and validity (p. 62) for a discussion of steps for ensuring good quality data after data collection.

References

Holaday, B., and A. Turner-Henson. "Response effects in surveys with school-age children." *Nursing Research*, vol. 38, no. 4, Jul.–Aug. 1989, pp. 248–250.

United Nations. "Household Income and Expenditure Surveys: A Technical Study." New York: National Household Survey Capability Programme, 1989.

De Leeuw, E.D., N. Borgers, and A. Smits. "Pretesting Questionnaires for Children and Adolescents." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. John Wiley & Sons, 2004, pp. 409–429. Available at https://dspace.library.uu.nl/bitstream/handle/1874/44538/EdL-pretesting+2004.pdf?sequence=1.

Exhibit III.1. Checklist for survey design and data collection

Survey instrument

1. Overall survey design

- ☐ Survey includes questions necessary to measure child labor: see the Checklist for Defining and Measuring Child Labor.
- ☐ Documentation identifies the sources of any survey questions that come from existing surveys, including whether and how they were adapted.
- ☐ Survey includes informed consent (for adults) or assent (for minors) language, either in survey instrument or in a separate consent/assent form.
- ☐ Question formats are easy to read and easy for enumerators to fill out.
- ☐ Response options are complete, including options for "don't know," "refuse," and "other," when appropriate.
- ☐ Questions specify allowable ranges for all questions (for example, days worked per week may not exceed seven).
- ☐ For questions with an "other, specify" option, the survey provides a place to enter a text response.
- ☐ Enumerator instructions are clear for all questions. Enumerators have received instruction about whether or not to read questions aloud, whether and how to assist respondents with difficult questions, and so on.
- ☐ Age questions include strategies for enumerators to use to estimate age accurately.
- ☐ Modules or questions targeted at children are appropriate for the target age.

2. Skip logic

- ☐ All skip logic instructions lead to the correct questions.
- ☐ Enumerator instructions are clear for all questions that require skips.

3. Document key components of translation protocol

- ☐ Describe how translation was conducted.
- ☐ Document all steps of the translation, including the original text, proposed translations, and final translations.
- ☐ Document any guides or tools used to facilitate nonwritten translation.

4. Identifying survey respondents

- ☐ Household and household head are clearly defined, and instructions clearly instruct enumerators to read the definitions to respondents.
- ☐ Each survey section clearly indicates who should respond, with a clear definition of the respondent.

- ☐ Sections about children indicate whether the child should answer or an adult should answer on the child's behalf.
- ☐ Sections targeted at children indicate what assistance from adults, if any, is allowable.

5. Instructions for absent respondents

Survey instructions indicate what to do if a targeted respondent is not available, including the following:

- ☐ How many attempts to make to reach the respondent
- ☐ How to follow-up to find an absent respondent, for example by phoning or returning at a specific time
- ☐ In what cases enumerators should interview someone else (a proxy) in the respondent's place
- ☐ Which persons they may interview as a proxy for an absent respondent.
- ☐ If enumerators should gather information on respondents' availability, such as contact phone numbers or times they are available, the survey instrument or a tracker document provides a place for enumerators to record this information.

Data collection protocol

6. Interviewer selection and training

- ☐ Enumerators are experienced and vetted.
- ☐ Enumerators are trained thoroughly on the specifics of the survey.
- ☐ Enumerators are tested on their ability to administer surveys correctly before hiring.
- ☐ Enumerator training includes how to conduct interviews with children, including where to conduct the interview, who may be present for the interview, and what enumerators should do if they discover child abuse or endangerment.
- ☐ As appropriate for the context, enumerators are assigned to interview respondents of the same gender.

7. Pre-testing and piloting

- ☐ Survey is pre-tested with adult respondents with similar education levels to the target adult population.
- ☐ Survey is pre-tested with child respondents with similar education levels to the target child population.
- ☐ Translations of questions on key topics and questions with possibly challenging language or terminology are pre-tested.

8. Monitoring data collection

- ☐ Field supervisors monitor response rates as the survey is in progress.
- ☐ Enumerators review each survey immediately after completion to check for completeness and errors in recording responses.

- ☐ Supervisors review surveys daily for completeness and logical consistency and follow up with enumerators to address errors.
- ☐ Supervisors call or visit some households to verify correct selection of household and to verify some responses.
- ☐ The grantee or contractor runs diagnostic tests on the data to look for errors such as incomplete surveys, implausible or impossible values, or other inconsistencies.
- ☐ Enumerators document information about how an interview was performed, including whether they used interpreters and/or proxy respondents, whether they estimated age information, and other details.

9. Reporting on data collection

- ☐ If the data collector is a contractor, that individual submits a report on data collection, including rates of nonresponse and reasons for nonresponse.

Exhibit III.2. Checklist for defining and measuring child labor

This checklist outlines the key considerations of defining child labor.

Developing definitions of child labor

1. Definitions specify the following key components:

- ☐ Age ranges
- ☐ Hours worked
- ☐ Activities that should be included in the term being defined

2. Definitions are measurable, allowing for any child to be categorized in the labor category being defined.

- ☐ References to age ranges are exhaustive and non-overlapping, covering all children ages 5 through 17.
- ☐ Maximum hours of work permitted without being considered child labor is specified per day and per week.
- ☐ Maximum hours of work permitted without being considered child labor is specified for all age ranges.

3. Definitions are consistent with international conventions on child labor.

- ☐ Age cutoffs associated with different ranges of allowable hours are consistent with those established in ILO Convention 138.
- ☐ Thresholds for allowable hours worked without being considered child labor are consistent with ILO Convention 138.
- ☐ Thresholds for hours of household chores are consistent with the International Conference on Labor Statistics guidance on children's participation in household chores.
- ☐ Every line or section of the definitions includes a citation of the legal framework (for example, an ILO convention) on which it is based.

4. Appropriate definition(s) are reported if national definitions differ from international definitions of child labor.

- ☐ If national definitions are <u>less strict</u> than international guidelines (for example, if they allow children to begin work at a younger age), reports include prevalence of child labor according to international definitions (in this case, reporting prevalence according to national definitions is optional).
- ☐ If national definitions are <u>more strict</u> than international guidelines, reports include prevalence of child labor according to both definitions.

Measuring child labor

5. All components of the definitions appear in surveys for all children ages 5 through 17, including the following, at a minimum:

- ☐ Child age, either through date of birth or age as of a certain date
- ☐ Number of hours worked per week or per day
- ☐ Time of day when working
- ☐ Activities performed (not limited to any key industry/sector of interest)
- ☐ Working conditions, including sufficient detail to identify hazardous or worst forms of labor

6. Survey questions put respondents at ease when asking sensitive questions, including questions about child labor.

- ☐ Survey instruments include statements that reassure respondents that their responses are confidential and they should answer honestly.
- ☐ Enumerator training covers strategies to put respondents at ease when they answer sensitive questions.
- ☐ If grantees do not expect respondents to answer honestly, surveys use list randomization to allow respondents to provide information without revealing their answer directly (see Karlan and Zinman [2012] reference on list randomization in the online resources list for more information on this approach).

Exhibit III.3. Annotated sample household survey on child labor

As discussed in the survey development section, using and adapting existing, internationally recognized surveys saves time and can improve survey quality when constructing new surveys. The International Labour Organization's (ILO) child labor survey (SIMPOC), is a helpful starting point when developing a new child labor survey. However, it is important to tailor survey questions to the local context.

This survey was modeled after the ILO's SIMPOC survey. We show only Part 3, which is administered to children. We have not modified the survey in any way other than to add text boxes to flag useful examples:

[✓] We point out useful survey strategies with green text boxes marked with a ✓.

[✓] We point out key data required for identifying child labor in blue boxes marked with a ✓.

[?] We point out items that may require customization with yellow text boxes marked with a ?.

Note that we have not marked everything you should review if using an existing survey as the basis for your own survey. Always review an existing survey carefully to identify things that should change to fit the local context. We hope that this sample questionnaire can serve as a useful tool for child labor researchers.

Child Labor Baseline Survey

This survey is structured in three sections. Part 1 and Part 2 are addressed to the most knowledgeable adult member of the household, whereas Part 3 is addressed directly to children (5 to 17).

1. **PART I: Adult questionnaire - Addressed to the most knowledgeable member of the household**
2. **PART II: Household characteristics - Addressed to the most knowledgeable member of the household**
3. **PART III: Child questionnaire - Addressed to every child (5–17) in the household**

[✓] These notes clearly indicate who responds to each section of the survey: the household head is clearly identified as the most knowledgeable member of the household and the children to be interviewed are correctly identified as all children ages 5 through 17. For sections about children, consider whether the most knowledgeable member should be the children's primary caregiver.

[?] Consider including a note specifying which children are considered part of the household, since it may not be obvious in the case of long-term guests, children staying somewhere else temporarily, or multi-household dwellings.

CHILD LABOUR SURVEY

(Addressed to the most knowledgeable member of the household)

GENERAL INFORMATION

Dept.....

Commune.....

Section Commun. .....

ADDRESS OF HOUSEHOLD..

PHONE NUMBER (----------

ENUMERATION AREA CODE..... (SDE)

HOUSE/STRUCTURE NO. |_|_|_|_|

HOUSEHOLD ID NUMBER |_|_|_|_|

[?] In some cases, enumerators must make several attempts to interview absent household members. Consider including space to record information on each interviewer visit, apart from the date of the final visit. The survey should also record the final interview's result, whether it was refused, completed, or otherwise. Refer to the ILO's SIMPOC survey for an example of how to record information on multiple visits.

Consider including space to record respondents' availability, such as dates and times available to facilitate revisits for household members who were absent during the first interview attempt.

INTERVIEWER VISIT

DATE/MONTH/YEAR : |_|

ELIGIBILITY

Number of persons in the household.....

Number of children (5-17) ..... (If 0, Thank the person and end the interview)

[?] Include a clear definition of the household and instructions for the enumerator to read the definition aloud to the person identified as the household head. Specify in the definition which children are considered part of the household. Provide instructions on what to do if any children (or other household members) are unavailable to participate in the interview. Options include interviewing an adult who is knowledgeable about the child or returning to the household later to attempt to interview the child. Include space for the enumerator to indicate if a child was unavailable and, if so, whether the enumerator interviewed a proxy respondent.

[✓] Section headers remind the enumerator about the correct respondent to interview for each section.

PART III CHILD QUESTIONNAIRE
Ask every child (5-17) in the household

SECTION IX — Educational Attainment of All Children (5-17)

							Skip to Question	
Serial No in A1								
Name of household member								
Age of household member							Children Aged 5-9 years	Children Aged 10-17 years
C1. Can you read and write a short, simple statement with understanding in any language? 1. Yes…… 2. No……			1 2	1 2				
C2. Are you attending school or pre-school during the current school year? 1. Yes…… 2. No……					1 2	1 2	→C3 →C8	
C3. What is the level of school and grade that you are currently attending? *Level: (L) Grade (G)* 1. Pre-school…… 2. Primary…… 3. Secondary……					L / G 1 2 3	L / G 1 2 3		
C4. At what age did you begin primary school? (*If C3=1 write 99*) *(Age in completed years…)*……					|__|__|	|__|__|		
C5. Did you miss any school day during the past week? 1. Yes…… 2. No……					1 2	1 2	→C6 →C17	→C6 →C14
C6. How many school days did you miss during the past week? *(Write the number of days)*……	|__|	|__|	|__|	|__|	|__|	|__|		
C7. Why did you miss school day(s) during the past week? *(Read each of the following options and circle two most appropriate option)* 1. School vacation period…… 2. Teacher was absent …… 3. Bad weather conditions…… 4. To help family business…… 5. To help at home with household tasks … 6. Working outside family business…. 7. Illness/ Injury/disablement …… 8. Other ……	7 8	7 8	7 8	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8	C17	C14
Other (Specify)……								

[✓] The survey uses serial numbers established for each household member in the first section of the survey to link to household members' information gathered in different sections of the survey.

[✓] *Key data for measuring child labor.* The survey collects age data on every child ages 5 through 17.

[✓] This column ensures that skip patterns are clear to the enumerator for each question.

[?] Given that the age of the child is fundamental to identifying child labor, use strategies for obtaining accurate age information if respondents do not know all children's dates of birth. Strategies include asking for the date of birth or anchoring the birth year on a major event. Include suggestions for using these strategies in the survey as a reminder (in addition to including it in enumerator training).

[?] Include instructions to enumerators to note if they have estimated a child's age (and enumerator training should provide best practices for estimating age).

[?] The numerical answer to question C6 should not exceed "7." To remind enumerators, consider specifying an allowable range for answers to such questions, for example, including "(No more than 7)" next to the response space.

Serial No in A1	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Skip to Question**	
Name of household member								
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Children Aged 5-9 years**	**Children Aged 10-17 years**
C8. Have you ever attended school?								
1. Yes……..	1	1	1			1	→C10	
2. No……	2	2	2			2	→C9	
C9. Why have you never attended school? (*Read each of the following options and circle the most appropriate option)*								
1. Too young ……..						1		
2. Disabled/ illness……				2	2	2		
3. No school/school too far……				3	3	3		
4. Cannot afford schooling……				4	4	4		
5. Family did not allow schooling…				5	5	5		
6. Not interested in school…….				6	6	6		
7. Education not considered valuable.				7	7	7	C17	C14
8. School not safe……...				8	8	8		
9. To learn a job……	9	9	9	9	9	9		
10. To work for pay ……	10	10	10	10	10	10		
11. To work as unpaid worker in family business/farm ……..								
12. Help at home with household tasks…	11	11	11	11	11	11		
13. Other ……..	12 13	12 13	12 13	12 13	12 13	12 13		
Other(Specify)								

C10. What is the highest level of school and grade you have attended?	L	G	L	G	L	G	L	G	L	G	L	G
Level (L) Grade (G)												
1. Pre-school……..……	1	|__|	1	|__|	1	|__|	1	|__|	1	|__|	1	|__|
2. Primary……	2	|__|	2	|__|	2	|__|	2	|__|	2	|__|	2	|__|
3. Secondary……	3	|__|	3	|__|	3	|__|	3	|__|	3	|__|	3	|__|
4. University ……	4	|__|	4	|__|	4	|__|	4	|__|	4	|__|	4	|__|
C11. At what age did you begin primary school? *(If C3=1 write 99) (Age in completed years)……..*	|__|__|		|__|__|		|__|__|		|__|__|		|__|__|		|__|__|	
C12. At what age did you leave school? *(Age in completed years)……..*	| | |		|__|__|		|__|__|		|__|__|		|__|__|		|__|__|	

[✓] Note that skip patterns here are logical. If children indicate they have never attended school, questions concerning their educational attainment are skipped. Correct skip patterns ensure logical question flow.

[✓] Question formats are easy to read: questions that the enumerators should ask are in boldface type, instructions to enumerators are in italic type, and possible responses are in spaces below the question. This design allows enumerators to quickly understand what to read aloud and what to fill out when conducting the interview.

[✓] The enumerator receives several prompts in various sections of the survey to input the age of the child being interviewed. This reminds enumerators of the correct skip patterns to follow, because skip patterns differ by the child's age.

[✓] Responses are coded as numerical values. This facilitates the use of the survey as a codebook for analysis later on. When creating the data set based on the survey data, ensure that variables in the final data set are recorded in the same way responses are coded in the survey.

[?] Some of these response options might be difficult for the youngest respondents. Consider rephrasing difficult phrases like "Cannot afford schooling," and "Education not considered valuable."

Serial No in A1							Skip to Question	
Name of household member								
Age of household member			|__|__|	|__|__|	|__|__|	|__|__|	Children Aged 5-9 years	Children Aged 10-17 years
C13. Why did you leave school? (*Circle the most appropriate option*)								
1. Completed his/her compulsory schooling (IF C10=**X**)	1	1				1	C17	
2. Too old for school…….	2	2				2		
3. Disabled/ illness……	3	3				3		
4. No school/school too far……	4	4				4		
5. Cannot afford schooling…	5	5				5		
6. Family did not allow schooling.	6	6				6		
7. Poor in studies/not interested in school.	7	7				7		
8. Education not considered valuable	8	8				8		
9. School not safe…….....	9	9				9		
10. To learn a job……	10	10				10		
11. To work for pay as employee or (as paid/ unpaid worker) in family business or farm.....	11	11	11	11	11	11		
12. Help at home with household tasks.....	12	12	12	12	12	12		
13. Other (Specify)…….	13	13	13	13	13	13		
Other(Specify)								
C14. Have you ever attended/are you currently attending a vocational / skills training course outside of school?								
1. Yes.....				1	1	1		→C15
2. No……...				2	2	2		→C17
C15. Have you /will you obtain a certificate for this vocational training?								
1. Yes	1	1	1	1	1	1		→C16
2. No……...	2	2	2	2	2	2		→C17
C16. Describe subject of vocational training received/being received. (e.g. Carpentry, Car repair, Nursing, etc.) *(If more than one then indicate the most important)*								

SECTION X	Current Economic Activities Status of All Children (5-17)						Skip to Question	
Serial No in A1	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|		
Name of household member								
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	Children Aged 5-9 years	Children Aged 10-17 years
Economic Activity								
C16.0 Have you been born here in the commune? 1. Yes…… 2. No……	1 2	1 2	1 2	1 2	1 2	1 2	If 1 → C17 If 2 → C16.1	
C16.1 If no, how many years since you are living here?	1 2 3 4 5 or+ 99=Don't know							
C16.3 Did someone else than your parents bring or send you here? 1. Yes…… 2. No……	1 2							
C 17. Did you engage in any work at least one hour during the past week? *(As employee, self-employed, employer or unpaid family worker)* 1. Yes…… 2. No……	1 2							
C18. During the past week, did you do any of the following activities, even for only one hour? *(Read each of the following questions until the first affirmative response is obtained)*	**1= YES 2=NO**							
(a) Run or do any kind of business, big or small, for himself/herself or with one or more partners? *Examples: Selling things, making things for sale, repairing things, guarding car, hairdressing, crèche business, taxi or other transport business, having a legal or medical practice, performing in public, having a public phone shop, barber, shoe shining etc.*	|__|			|__|	|__|	|__|		
(b) Do any work for a wage, salary, commission or any payment in kind (excl. domestic work)? *Examples: a regular job, contract, casual or piece work for pay, work in exchange for food or housing.*	|__|	|__|	|__|	|__|	|__|	|__|	If any "YES" → C20	
(c) Do any work as a domestic worker for a wage, salary or any payment in kind?	|__|	|__|	|__|	|__|	|__|	|__|	Otherwise	
(d) Help unpaid in a household business of any kind? *(Don't count normal housework.)*	|__|	|__|	|__|	|__|	|__|	|__|		

[?] Respondents might have lived there for less than one year. Consider adding an option for less than one year, and an indication of when to round up or down to a whole year for those who have lived there for a part of a year (for example, for two and a half years).

[?] Respondents also might not know how long they have lived in the commune, so it is a good idea to include "don't know" as an option. For sensitive questions, it may be worth including the response "refuse" as well (and respondents always have the right to refuse to answer specific questions). When including "refused" or "don't know" responses, be sure to avoid using numeric responses as these can lead to errors when using the data to calculate averages or other statistics. Software packages like STATA allow you to code such responses as missing values with letters like ".d" or ".r".

[?] Very young respondents may not be able to say how many years they have lived in their commune. You may want to offer alternative phrasing, such as, "How old were you when you moved to the commune?"

[✓] ***Key data for measuring child labor.*** The C18 series of questions collects data on children's work activities.

Examples: Help to sell things, make things for sale or exchange, doing the accounts, cleaning up for the business, etc.						
(e) Do any work on his/her own or the household's plot, farm, food garden, or help in growing farm produce or in looking after animals for the household? *Examples: ploughing, harvesting, looking after livestock.*	|__|	|__|	|__|	|__|	|__|	|__|
(f) Do any construction or major repair work on his/her own home, plot, or business or those of the household?	|__|	|__|	|__|	|__|	|__|	|__|
(g) Catch any fish, prawns, shells, wild animals or other food for sale or household food?	|__|	|__|	|__|	|__|	|__|	|__|
(h) Fetch water or collect firewood for household use?	|__|	|__|	|__|	|__|	|__|	|__|
(i) Produce any other good for this household use? *Examples: clothing, furniture, clay pots, etc.*	|__|	|__|	|__|	|__|	|__|	|__|

Serial No in A1	| | |	| | |	| | |	| | |	| | |	| | |	Skip to Question	
Name of household member								
Age of household member	| | |	| | |	| | |	| | |	| | |	| | |	Children Aged 5-9 years	Children Aged 10-17 years
C19. Even though you did not do any of these activities in the past week, do you have a job, business, or other economic or farming activity that you will definitely return to? *(For agricultural activities, the off season in agriculture is not a temporary absence).* 1. Yes…… 2. No……	1 2	1 2	1 2	1 2	1 2	1 2	→C20 →C31	
C20. Describe the main job/task you were performing e.g. carrying bricks; mixing baking flour; harvesting maize; etc. *("Main" refers to the work on which (NAME) spent most of the time during the week.)*								
Job/Task								
C21. Describe briefly the main activity i.e. goods produced and services rendered where you are doing this job or task								
Activity / Type								

[✓] *Key data for measuring child labor.* Question C20 gathers additional data on the work activities children performed.

[✓] As these skip patterns illustrate, the enumerator asks different sets of questions according to the age of the children. Younger children are not asked detailed questions about work and pay. Carefully target different age groups with modules that are appropriate to their age level.

Serial No in A1							Skip to Question
Name of household member							
Age of household member							Children Aged 10-17 years
C22. In addition to your main work, did you do any other work during the past week?							
1. Yes……			1	1	1	1	
2. No……			2	2	2	2	
C23. For each day worked during the past week how many hours did you actually work? *Main (M) Other (O)*	M	O	M	O	M	O	
1. Monday……							
2. Tuesday……							
3. Wednesday……							
4. Thursday……							
5. Friday……							
6. Saturday……							
7. Sunday……							
TOTAL							
C24. During the past week when did you usually carry out these activities?							
For ALL children (including children attending school):							
1. During the day (between 6 a.m. and 6 p.m) …..	1	1	1	1	1	1	
2. In the evening or at night (after 6 p.m.)	2	2	2	2	2	2	
3. During both the day and the evening (for the entire day).	3	3	3	3	3	3	
4. On the week-end….....	4	4	4	4	4	4	
5. Sometimes during the day, sometimes in the evening	5	5	5	5	5	5	
ADDITIONAL: ***For children attending school ONLY (If C2=YES):***							
6. After school.....	6	6	6	6	6	6	
7. Before school.....	7	7	7	7	7	7	
8. Both before or after school…..…..	8	8	8	8	8	8	
9. On the week-end……..	9	9	9	9	9	9	
10. During missed school hours/days.....	10	10	10	10	10	10	

[✓] ***Key data for identifying child labor.*** **The survey gathers information on the number of hours worked per day for every day in the past week.**

[?] Pilot the survey to see if the youngest respondents are able to say how many hours they worked each day. If they have trouble, consider using a visual aid to help them answer.

[✓] The numerical answer to this question cannot exceed "24." Consider adding two reminders here: (1) that the number cannot exceed 24 and (2) to double-check extreme values, such as values over 16 per day.

[✓] ***Key data for identifying child labor.*** **The survey collects data on the time of day when children worked in the past week.**

Serial No in A1	|_|_|_|	|_|_|_|	|_|_|_|	|_|_|_|	|_|_|_|	|_|_|_|	**Skip to Question**
Name of household member							
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Children Aged 10-17 years**
C25. Where did you carry out your main work during the past week?							
1. At (his/her) family dwelling…	1	1	1	1	1	1	
2. Client's place ……	2	2	2	2	2	2	
3. Formal office ……..	3	3	3	3	3	3	
4. Factory / Atelier ……	4	4	4	4	4	4	
5. Plantations / farm / garden…….	5	5	5	5	5	5	
6. Construction sites……..	6	6	6	6	6	6	
7. Mine / quarry……..	7	7	7	7	7	7	
8. Shop / kiosk / coffee house / restaurant / hotel	8	8	8	8	8	8	
9. Different places (mobile)……..	9	9	9	9	9	9	
10. Fixed, street or market stall	10	10	10	10	10	10	
11. Pond / lake / river.....	11	11	11	11	11	11	
12. Other……	12	12	12	12	12	12	
Other (specify)							
C26. For your main job/work were you a/an….?							
1. Employee…….....	1	1	1	1	1	1	→C27
2. Own account worker (His/her own business without employees)					2	2	
3. Employer (His/her own business with employees)							
4. Member of producers` cooperatives					3	3	C28
5. Unpaid family worker….					4	4	
					5	5	
C27. What was the mode of payment for the last payment period?							
1. Piece rate.....							
2. Hourly.....	1	1	1	1	1	1	
3. Daily.....	2	2	2	2	2	2	
4. Weekly.....	3	3	3	3	3	3	
5. Monthly.....	4	4	4	4	4	4	
6. Upon completion of task.....	5	5	5	5	5	5	
7. Other (specify).....			6	6	6	6	
			7	7	7	7	
Other (specify)							

[?] Technical terminology such as "piece rate," "overtime pay," and "accident insurance," might be difficult for children to understand. Find terminology that child respondents will understand. Do not include questions about concepts that child respondents are unlikely to understand as that might distress the children and generate inaccurate data.

[?] The space for recording answers to "Other (specify)" may be too narrow here. Electronic data collection can solve space constraints, but on paper questionnaires, this can be improved by incorporating more lines and expanding the horizontal space for enumerators to write responses. This comment applies to other questions as well that allow "Other (specify)" as a response.

Serial No in A1	| | |	| | |	| | |	| | |	| | |	| | |	Skip to Question	
Name of household member								
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	Children Aged 5-9 years	Children Aged 10-17 years
C28. What is your average monthly income from the main work?	| | | |	| | | |	| | | |	| | | |	| | | |	| | | |		
C29. What do you usually do with your earnings? *(Multiple answers are allowed)*								
1. Give all/part of money to my parents	1	1	1	1	1	1		
2. Employer gives all/part of money to my parents	2	2	2	2	2	2		
3. Give all/part of money to the person who bring or send me here	3	3	3	3	3	3		
4. Employers gives all/part of money to the person who bring or send me here	4	4	4	4	4	4		
5. Pay my school fees	5	5	5	5	5	5		
6. Buy things for school	6	6	6	6	6	6		
7. Buy things for household	7	7	7	7	7	7		
8. Buy things for myself	8	8	8	8	8	8		
9. Save	9	9	9	9	9	9		
10. Other.....	10	10	10	10	10	10		
Other (specify)								
C30. Why do you work?								
1. Supplement family income…								
2. Help pay family debt……								
3. Help in household enterprise…								
4. Learn skills……								
5. Schooling not useful for future..……								C33
6. School too far / no school …….								
7. Cannot afford school fees……								
8. Not interested in school……								
9. To temporarily replace someone unable to work.								
A. Job Search								
C31. Were you seeking work during the last week?								
1. Yes……..……..	1	1	1	1	1	1		
2. No……..……	2	2	2	2	2	2		
C32. At any time during the past 12 months did you engage in any work?								
1. Yes……..……..	1	1	1	1	1	1	→C33	
2. No……..……	2	2	2	2	2	2	→C41	

[?] For question like C30 that have multiple response options, indicate whether respondents may select multiple responses.

[?] These response options may be difficult for young respondents to understand. Rather than specifying that they are paying off debt or that they cannot afford school fees, they may say something like, "We need the money." Pilot the survey to see how well young respondents understand these options. If they do not understand them well, consider combining response options and using simpler terms.

SECTION XI	Health and Safety Issues about working children (5-17)						
Serial No in A1	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Skip to Question**
Name of household member							
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Children Aged 5-9 years** / **Children Aged 10-17 years**
C33. Did you have any of the following in the past 12 months because of your work? (*Read each of the following options and mark "YES" or "NO" for all options*)	**Y= YES N=NO**	**Y= YES N=NO**	**Y= YES N=NO**	**Y= YES N=NO**	**Y= YES N=NO**	**Y= YES N=NO**	If all **"NO"** → **C36** **Otherwise → C34**
1. Superficial injuries or open wounds	1|___|	1|___|	1|___|	1|___|	1|___|	1|___|	
2. Fractures……..	2|___|	2|___|	2|___|	2|___|	2|___|	2|___|	
3. Dislocations, sprains or stains…					3|___|	3|___|	
4. Burns, corrosions, scalds or frostbite					4|___|	4|___|	
5. Breathing problems.....					5|___|	5|___|	
6. Eye problems.....					6|___|	6|___|	
7. Skin problems…					7|___|	7|___|	
8. Stomach problems / diarrhea …					8|___|	8|___|	
9. Fever……..					9|___|	9|___|	
10. Extreme fatigue.....					10|___|	10|___|	
11. Other (specify).....					11|___|	11|___|	
Other (specify)							
C34. Think about your most serious illness/injury, how did this/these affect your work/schooling?							
1. Not serious- did not stop work/schooling.	1	1	1	1	1	1	
2. Stopped work or school for a short time	2	2	2	2	2	2	
3. Stopped work or school completely.	3	3	3	3	3	3	
C35. Think about your most serious illness/injury, what were you doing when this happened?							
Job/Task							

[?] Respondents as young as 5 years old may not understand some of these response options, such as "superficial injuries or open wounds," "dislocations, sprains, or stains," and "burns, corrosions, scalds, or frostbite." Pilot the survey to determine how well young children understand these terms. If they do not understand them well, consider using simpler terms or training enumerators to explain the concepts to young children.

Serial No in A1	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Skip to Question**	
Name of household member								
Age of household member	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	**Children Aged 5-9 years**	**Children Aged 10-17 years**
C36. Do you carry heavy loads at work? 1. Yes..... 2. No.....	1 2	1 2	[illegible]	[illegible]	[illegible]	1 2		
C37. Do you operate any machinery/heavy equipment at work? 1. Yes…….. 2. No……	1 2	1 2	2	2	2	1 2	→ C38 → C39	
C38.What type of tools, equipment or machines do you use at work? (*Write down 2 mostly used*)	1…… 2……	1…… 2……	1…… 2……	1…… 2……	1…… 2……	1…… 2……		
C39. Are you exposed to any of the following at work? (*Read each of the following options and mark "YES" or "NO" for all options*)	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**		
1. Dust, fumes,	1|__|	1|__|	1|__|	1|__|	1|__|	1|__|		
2. Fire, gas, flames.....	2|__|	2|__|	2|__|	2|__|	2|__|	2|__|		
3. Loud noise or vibration.....	3|__|	3|__|	3|__|	3|__|	3|__|	3|__|		
4. Extreme cold or heat	4|__|	[illegible]	[illegible]	[illegible]	4|__|	4|__|		
5. Dangerous tools (knives etc).....	5|__|				[illegible]	5|__|		
6. Work underground.....	6|__|				[illegible]	6|__|		
7. Work at heights.....	7|__|				[illegible]	7|__|		
8. Work in water/lake/pond/river…….	8|__|				[illegible]	8|__|		
9. Workplace too dark or confined	9|__|				[illegible]	9|__|		
10. Insufficient ventilation.....	10|__|				[illegible]	10|__|		
11. Chemicals (pesticides, glues, etc.)..	11|__|				[illegible]	11|__|		
12. Explosives.....	12|__|	12|__|	12|__|	12|__|	12|__|	12|__|		
13. Other things, processes or conditions bad for your health or safety (specify).....	13|__|	13|__|	13|__|	13|__|	13|__|	13|__|		
Other (specify)								
C40. Have you ever been subject to the following at work? (*Read each of the following options and mark "YES" or "NO" for all options*)	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**	**1= YES 2=NO**		
1. Constantly shouted at …….	1|__|				[illegible]	1|__|		
2. Repeatedly insulted……..	2|__|				[illegible]	2|__|		
3. Beaten /physically hurt…	3|__|				[illegible]	3|__|		
4. Sexually abused (touched or done things to you that you did not want)	4|__|				[illegible]	4|__|		
5. Other (Specify)……	5|__|				[illegible]	5|__|		
Other (specify)								

[✓] ***Key data for identifying child labor.*** **This series of questions (C36–C40) covers working conditions to identify hazardous child labor.**

[?] Pilot the survey to determine whether children understand these terms. If not, consider using simpler terms or training enumerators on ways to explain them in simple terms.

[?] Consider whether or not you should ask questions about the worst forms of child labor of any children, giving special consideration to young children.

SECTION XII	Household Tasks of Children (5-17)						Skip to Question	
Serial No in A1	| | |	| | |	| | |	| | |	| | |	| | |		
Name of household member								
Age of household member	| |__|	| |__|	| |__|	| |__|	| |__|	| |__|	Children Aged 5-9 years	Children Aged 10-17 years
C41. During the past week did you do any of the tasks indicated below for this household? (*Read each of the following options and mark "YES" "NO" for all options*)	Y= YES N=NO	Y= YES N=NO	Y= YES N=NO	Y= YES N=NO	Y= YES N=NO	Y= YES N=NO	If any "**YES**" →**C42** Otherwise **END** for this HH member. Go to the next child in Section II.	
1. Shopping for household....	1| |	1| |	1| |	1| |	1| |	1| |		
2. Repair any household equipment …			2| |	2| |	2| |	2| |		
3. Cooking.....			3| |	3| |	3| |	3| |		
4. Cleaning/house.....			4| |	4| |	4| |	4| |		
5. Washing clothes.....			5| |	5| |	5| |	5| |		
6. Caring for children/old/sick.....			6| |	6| |	6| |	6| |		
7. Other household tasks.....			7| |	7| |	7| |	7| |		
Other (Specify)								
C42. During each day of the past week how many hours did you do such household tasks? (Record for each day separately)								
1. Monday……			|__|__|	|__|__|	|__|__|	|__|__|		
2. Tuesday…….			|__|__|	|__|__|	|__|__|	|__|__|		
3. Wednesday…….			|__|__|	|__|__|	|__|__|	|__|__|		
4. Thursday…….			|__|__|	|__|__|	|__|__|	|__|__|		
5. Friday……..	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|		
6. Saturday……..	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|		
7. Sunday…….	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|	|__|__|		
TOTAL	| | |	| | |	| | |	| | |	| | |	| | |		
C43. During the past week when did you usually carry out these activities? ***For ALL children (including children attending school):***								
1. During the day (between 6 a.m. and 6 p.m.) …..	1	1	1	1	1	1		
2. In the evening or at night (after 6 p.m.)	2	2	2	2	2	2		
3. During both the day and the evening (for the entire day).	3	3	3	3	3	3		
4. On the week-end….....….	4	4	4	4	4	4		
5. Sometimes during the day, sometimes in the evening	5	5	5	5	5	5		
ADDITIONAL: ***For children attending school ONLY (If C2=YES):***								
6. After school.....	6	6	6	6	6	6		
7. Before school.....	7	7	7	7	7	7		
8. Both before or after school…..…..	8	8	8	8	8	8		
9. On the week-end……..	9	9	9	9	9	9		
10. During missed school hours/days.....	10	10	10	10	10	10		

[?] The numerical answer to this question should not exceed "24"; this should be noted for the enumerator.

[?] Young children may have a hard time specifying how many hours they worked. Consider offering a visual aid to help.

C44. Has (NAME) been interviewed in the company of an adult or an older child? *1. Yes* *2. No*	1 2	1 2	1 2	1 2	1 2	1 2	**END** for this HH member. Go to the next child in Section II.

[✓] The survey records information on whether an interview with a child occurred in the company of other members of the household. This is essential information to record because the presence of adults may influence a child's answer to child labor questions.

[?] Whether the adults were able to hear the child's responses is also important; consider recording that information as well.

IV. DATA FORMAT AND VALIDITY

This chapter covers formatting and validating data, including coding data to simplify analysis, being consistent with response codes in corresponding surveys, and appropriately dealing with missing data and impossible or implausible values.

A. Document changes to the data

Preserve the original versions of all data files. The original data provide an important record in the data cleaning and validation process. Original data files can also be checked to investigate errors made in the cleaning process.

Identify and correct errors in the data. Look for errors such as skipped questions, implausible or impossible values, or responses that are inconsistent with the respondent's answers to other questions. If the budget permits, use double data entry to minimize data-entry errors when data are collected using paper forms—that is, have two people enter the data from each completed survey independently, and check for instances in which the two entries do not agree. Whether using an electronic instrument or a paper survey, develop rules for correcting errors made during the data collection process, including rules and procedures for the types of errors you will attempt to correct and the types of corrections you will make. The first step is often contacting the data collector to verify any data you suspect may be incorrect. The University of Michigan's *Cross-Cultural Survey Guidelines* chapter on data processing and statistical adjustment provides more detail on data cleaning.

Keep a log of any changes to the data. Cleaning and validating data typically involves making some changes to the original survey data to correct errors as well as to construct new variables: document all such changes. Write a code (such as ".do file" in STATA) to examine and correct your survey data consistently and more thoroughly than is possible by manually looking at and fixing the data. To do this, use the software to document the written program that executes the changes to the data. Clearly document any additional manual adjustments made and, if possible, include such adjustments in the program so the cleaning can be replicated. If no program was used, document clearly how you examined the data and what specific changes were made. Complete cleaning of data before creating new variables (constructs) to ensure that new variables are calculated based on accurate data. Submit documentation of changes made to the data along with the public use version of the data files.

B. Produce a codebook to allow data users to interpret the data

A codebook is a guide that describes all of the variables in the data set. If possible, develop the codebook as survey instruments are being developed, to ensure consistency and completeness. At the latest, prepare the codebook before data entry to ensure that data are entered consistently. A codebook typically contains the following elements for each variable:

- Variable name: the exact name of each variable in the data set.
- Variable label: a description of the variable.

- Question text: the text from the survey instrument. This text may be used as the variable label.
- Values: the coded values in the data (for example, 0, 1), including values for missing data, where relevant.
- Value labels: the description of the codes ("no" and "yes").
- Skip patterns: instructions to skip questions that do not apply based on answers to previous questions. Where applicable, indicate who should have responded to the question.

Ensure that codebooks match the survey questions, response codes, and data set value labels for each variable. Codebooks can serve as a tool to check the consistency of the labels as well as a helpful document to guide the research of external analysts.

Include in the codebook any variables such as weights that are in the data set but not in the survey. If your analysis requires the use of weights, your data set will use weights. If this is the case, be sure to label the weights variables and include them in the codebook. Other such variables used for analysis but not in the survey include strata IDs, PSU IDs, enumerator IDs, and household IDs, among others.

C. Code data consistently with the final survey instruments

Label all variables and response values in a data set consistently with the way the responses are coded in your survey. For example, if the survey specifies that enumerators enter 1 for "yes" and 0 for "no," use the same codes in the data set. Consistent labels make it easy to accurately interpret the data and relate the survey to the data.

D. Format data set and variables in a way that is conducive to analysis

Code and format data sets to improve usability. Coding of variables refers to the process of storing survey response options as data in a data set. This process involves choosing a format and may involve assigning values. Numeric variables take on only numeric values and can be used in quantitative analysis. String variables may have numeric values, text values, or a combination of the two, and cannot be used easily for quantitative analysis. If response options were not assigned codes in the survey, add numeric codes in the data set and use the text responses as labels. If you must code your data in a way that differs from your survey to facilitate analysis, document such changes in your codebook. Data coding and formats determine how easy it is to use a data set. The following strategies improve usability.

Structure the data sets so they are easy to use. Consider creating two data sets—one at the household level, with household variables, and one at the child level, with information about individual children within the household. Under such an approach, the household data set would have one observation (row) per household, whereas the child-level data set would have one row for each child. Include a unique household ID in each data set to link the two data sets and allow

household-level variables to be included in child-level analyses. An example of household- and child-level data sets is shown in Table IV.1.

Table IV.1. Example of household and child data set structures

Household level		Child level		
Household ID	Number of children	Household ID	Child ID	Child's age
1	3	1	1	13
2	1	1	2	11
3	2	1	3	7
		2	1	4
		3	1	6
		3	2	2

Format categorical variables as numeric variables rather than string variables (variables stored as text). Categorical variables are variables that have a set number of response options. For example, a variable that records the time of day an activity occurred might have the response options "morning," "afternoon," or "evening." Code categorical variables as numeric variables because those are easier to use in analysis than string variables and less prone to errors if data are converted between formats. Include text labels for numeric values to facilitate analysis and interpretation. For the variable for time of day, for example, code responses as 1 (with label "morning"), 2 (with label "afternoon"), or 3 (with label "evening"). STATA's help site provides instructions on attaching labels to numeric values. If the survey has multiple-choice questions for which a respondent may select multiple options, create a new binary variable for each response option to make these variables easy to analyze. Be sure to label the variables clearly.

Table IV.2. Example of variables with variable labels, response values, and response value labels

Variable	Variable label	Response values	Response value labels
female	Respondent is female	0 1	Male Female
floor_type	Main type of floor in respondent's home	1 2 3	Dirt Wood Tile

Source: Sample variable descriptions.

For variables with two possible response options ("binary variables"), code responses as 0 or 1, with labels. For example, a gender variable could be coded 0 for male and 1 for female. Coding the variables in this way allows the user to quickly calculate the percentage of respondents in the category coded as 1; the mean of the binary variable is equal to the percentage in that category.

When coding female as 1, the mean of the gender variable would be equal to the percentage of female respondents. Code the category you are more likely to report (for example, children working in child labor) as 1. Coding variables as 0 or 1 also facilitates the analysis of regression coefficients, if used.

Review and consider recoding textual comments respondents make when they select "other, specify" to multiple choice questions. When survey instruments provide space for enumerators to enter "other, specify" responses, enter such responses into the data word for word. Review such responses as part of the data cleaning and validation process. If "other" responses fit into a question's existing categories, change the response to that existing category as part of the cleaning process; this process is known as "back-coding." Sometimes the review process reveals that many respondents gave similar "other" responses. When that occurs, group those responses in a new category with its own code and label. If any of these changes are made to the original data, clearly document such changes in the cleaning code or other documentation, as discussed above.

E. Code missing data logically and consistently

Code missing data values in a way that permits the analyst to identify the data as missing. When missing values are coded correctly, software programs will correctly and automatically exclude them from the calculation of means or other statistics. The following strategies make it easy to identify and, when appropriate, exclude missing data from analyses.

Code missing data as a non-number and label it as missing in the value label and codebook. When missing data are coded as a number, such as 999, -9, or 0, the statistical software cannot tell these are missing data and will calculate means incorrectly with no error warning. STATA and other software programs allow numeric variables to include nonnumeric values such as ".m" to indicate missing data. STATA and other programs also permit multiple missing codes to identify different types of data, as we describe in detail below.

Use codes to identify different types of missing data. STATA and other software programs allow for a range of different nonnumeric values, permitting the analyst to differentiate between data that are logically missing because of a skip pattern and data that are missing because of nonresponse. For example, in a question about someone's current grade level, people who are not in school will logically have a missing response because they were not asked the question if the skip pattern was followed correctly. Conversely, people who are in school might be missing a response if they do not answer the question or do not know. Use data codes to distinguish between these types of cases, as well as from other types of missing data, such as refusals. For instance, ".m" could indicate a logical skip, ".n" could indicate refusal to answer, and ".d" could indicate the respondent did not know the answer. It is helpful to use the same set of missing data codes across all variables in a survey. For example, if the code ".m" is selected to indicate a logical skip, use it to code logical skips wherever relevant, rather than ".m" for some logical skips and ".n" for others. Document how you code missing data in the codebook.

Label missing data codes to explain why the data are missing. Labels such as "don't know" or "refused" help analysts understand why responses are missing and help researchers modify future surveys to improve item response rates.

F. Address impossible or implausible values in the data

Review data for impossible values (such as someone working 36 hours in the last day) or implausible values (such as someone dropping out of school at age 2). Although careful design and implementation of surveys (including the training of enumerators) minimizes such values, develop a strategy for addressing any discovered during data analysis.

Define a cutoff for impossible or implausible values for all relevant variables. Impossible values are typically easy to define, but implausible values require more thought. For example, working 36 hours in a day is impossible, whereas working 22 hours is implausible but not impossible. Choose a cutoff beyond which values will be considered implausible.

Check the data consistently for impossible or implausible values. If you have predetermined a cutoff beyond which you consider responses to be impossible or implausible, you can check for all responses beyond these cutoff values. The "count" command in STATA is an easy way to do this. You can write checks into a program file to automate checking for such values.

Try to correct or validate impossible or implausible values in the data. Contact enumerators or their supervisors to ask about impossible or implausible values. They might be able to provide the correct response or an explanation for a valid but implausible response. If a correction is made, document it clearly, including an explanation for the change. The explanation may be included as a comment in the .do file (if using STATA, or in a similar file if using other software).

Develop a consistent rule for coding impossible or implausible values. If the field team does not have additional information, consider coding the impossible or implausible values as missing (and using a specific missing code) or, in some cases, replacing the impossible or implausible value with the closest plausible or possible value; this is known as "top-coding," It is best to recode variables that have impossibly or implausibly high values as missing (using a code that indicates that the value was set to missing during editing) unless you have strong reason to believe the true number is actually very high, such as when responses from the same household are consistent with the value in question being high. For example, an income variable may be implausibly high, but responses to other questions may be consistent with an unusually high income. In this situation, top-coding may be appropriate. If top-coding, replace data with the highest plausible value observed in the data. Do not use top-coding for more than 1 percent of cases. When in doubt, the safer approach is to recode a response as missing.

Exhibit IV.1. Checklist for data format and validity

Please review this entire checklist before beginning work with the data.

Documentation and codebook

1. All original files are preserved.

- ☐ Original and cleaned files are appropriately labeled.
- ☐ Program files that created the corrected files from the original files are clearly identified, if applicable.

2. All changes to the data are documented.

- ☐ All changes to the data are recorded in a separate program (for example, a .do file in STATA).
- ☐ The program file or other tracker document includes explanations for any corrections made.
- ☐ Any weights are clearly labeled in your data set.

3. Each data set's codebook contains the following elements:

- ☐ Variable name (the exact name of each variable in the data set).
- ☐ Variable label (a description of the variable).
- ☐ Question text (the text from the survey instrument). This text may be used as the variable label if it is not too long.
- ☐ Values (response codes used in the data (for example, 0, 1), including values for missing data where relevant).
- ☐ Value labels (the description of the codes; for example, "no" and "yes").
- ☐ Skip patterns (where applicable, a description of who should have responded to the question).
- ☐ Documentation of any variables that were constructed from the clean data, including a description of how they were constructed and from which variables in the clean data.

Data entry

4. Data are coded consistently with survey instruments.

- ☐ Variable names are consistent with survey questions.
- ☐ Variable value codes are consistent with survey instrument.
- ☐ Values labels are consistent with response options.

5. Data format simplifies analysis.

- ☐ Categorical variables are formatted as numeric variables.
- ☐ Binary variables are coded as 0 or 1.
- ☐ All values are labeled.

- ☐ Responses to "other, specify" questions are reviewed and possibly recoded.

6. Missing data are coded logically and consistently.

- ☐ Missing data codes are nonnumeric.
- ☐ Different types of missing data (for example, irrelevant questions, refusal to respond) have different codes and labels.
- ☐ Missing data labels indicate reason for being missing.
- ☐ Coding scheme for missing data is consistent throughout the data set.

7. Impossible or implausible values are addressed.

- ☐ A cutoff value for each continuous variable determines if a value is considered implausible or impossible.
- ☐ A consistent rule for coding impossible or implausible values is followed for all variables.
- ☐ Impossible or implausible values are corrected when possible.

V. ANALYSIS AND REPORTING

This chapter summarizes how to write baseline and follow-up reports, including the structure and topics of reports, templates for presenting results on the prevalence of child labor in baseline and follow-up reports, information on making tables and figures clear and complete, how to present statistical analysis, and how to present the discussion of findings.

A. Structure and topics for reports on prevalence of child labor

Include the following sections in baseline and follow-up reports on the prevalence of child labor:

1. Front sections

- ☐ Table of contents
- ☐ List of tables
- ☐ List of figures
- ☐ List of acronyms
- ☐ Executive summary

2. Background

- ☐ Country context
- ☐ Description of the intervention
- ☐ Project locations, if report is related to the implementation of a child labor-related project
- ☐ Research questions

3. Data collection and analysis

- ☐ Sampling design and sample size
- ☐ Survey instrument design and development
- ☐ Enumerator training
- ☐ Field work
- ☐ Data entry
- ☐ Data analysis
- ☐ Limitations of the study

4. Child labor estimates

- ☐ Prevalence of child labor
- ☐ Prevalence of children at risk of child labor
- ☐ Prevalence of hazardous child labor
- ☐ Prevalence of the worst forms of child labor (if applicable)

5. Conclusion and recommendations

- ☐ Summary of findings (for follow-up or endline reports written after a child labor project has been implemented, this should include a comparison between baseline and endline child labor prevalence rates)
- ☐ Implications for current or future program design and implementation, as appropriate

6. Appendices

- ☐ References
- ☐ Definitions of key terms
- ☐ Secondary results
- ☐ Survey instruments
- ☐ Training content
- ☐ Piloting process and results

Include child labor prevalence rates for the full sample and subgroups in the main body of the report, along with key results that describe the sample or that inform program design and strategy. For example, if the report is related to a project with an educational component, include analysis on school enrollment and attendance. For reports on projects with an awareness-raising component, providing analysis of household members' knowledge, attitudes, and practices related to child labor in the report because this information may make it easier to target the program. It is not necessary to include all data collected in the main body of the report. To keep the focus on the main findings, place less important tables, figures, and other information in an appendix or omit them. For example, in most cases, children's enrollment in school is more important than the number of classrooms in the closest school.

In the section on limitations of the study, describe any issues that may influence the interpretation of results. Many potential limitations come from issues that arise during data collection. For example, if data collection takes place during an intense political campaign, respondents may be less willing to respond to a survey, potentially reducing response rates. In cases where respondents have been asked to respond to many surveys, they may experience "survey fatigue," which could affect response rates or the quality of responses.

In the section on data entry, identify the program used to enter survey responses into a database as well as the quality control measures implemented to ensure high quality data entry. This includes indicating whether double data entry was employed, describing how missing data were coded, and explaining how impossible or implausible survey responses were handled. We discuss a number of potential quality control measures in the chapter on data format and validity.

In the data analysis section, indicate if statistical *t*-tests, regression analysis, or other methods were used, including whether and how weights were incorporated in the estimation of prevalence rates of child labor. The section below provides a few suggestions for providing a comprehensive analysis.

B. Presenting estimates of the prevalence of child labor

One primary purpose of a report on a child labor prevalence survey is to present the estimated prevalence of child labor, children at risk of child labor, hazardous child labor, and worst forms of child labor (if applicable) in the regions being studied. The goal of a follow-up report on the prevalence of child labor after an intervention has taken place is to show how these measures change over time.

The basic table template below provides an example of how to report results (Table V.1). In results tables, include estimated prevalence rates, sample sizes, standard errors, and 95 percent confidence intervals for all child labor indicators—that is, present the same tables for the prevalence rates for child labor, children at risk of child labor, hazardous child labor, and worst forms of child labor (if applicable). In a footnote, indicate the data source, how sampling weights were incorporated in estimates (if used), and other notes when necessary. (See Chapter II on sampling (p. 7) for more detail on when to use sampling weights.)

Table V.1. Example: Estimated prevalence of child labor in project regions

	Prevalence rate (%)	Standard error	95% confidence interval	n
All	59.3	0.9	57.5–61.1	2,422
By gender				
Girls	51.8	1.0	49.8–53.8	1,192
Boys	66.6	1.4	63.8–69.4	1,230
By age				
5–10 years old	39.8	1.2	37.4–42.2	1,112
11–15 years old	68.2	1.1	66.0–70.4	780
16–17 years old	87.1	2.3	82.6–91.6	530

Source: List source and other notes here, including information on how sampling weights were applied to generate estimates.

It is possible to tailor this basic template to accommodate project needs. For example, researchers may expand the table to include prevalence rates for other subgroups, such as urban/rural, or prevalence rates disaggregated further by region and age combinations (that is, prevalence rates for children of different ages within regions).

For follow-up reports, researchers can modify the table to include additional before-and-after columns to facilitate the comparison of prevalence rates between baseline and follow-up, as in Table V.2, below. To facilitate the analysis of changes in child labor prevalence rates over time, researchers may present the *p*-value associated with the *t*-tests comparing prevalence rates at baseline and follow-up (discussed in more detail later in this chapter).

Table V.2. Example: Estimated prevalence of child labor in project regions at baseline and follow-up

	Baseline Prevalence rate (%) (standard error)	Follow-up Prevalence rate (%) (standard error)	*p*-value of difference between baseline and follow-up	Baseline 95% confidence interval	Follow-up 95% confidence interval	Baseline n	Follow-up n
Overall	59.3 (0.9)	52.8 (1.1)	0.000	57.5–61.1	50.7–54.0	2,422	2,180
By gender							
Girls	51.8 (1.0)	52.0 (1.2)	0.890	49.8–53.8	49.7–54.3	1,192	1,073
Boys	66.6 (1.4)	53.6 (1.6)	0.000	63.8–69.4	50.5–56.7	1,230	1,107
By age							
5–10 years old	39.8 (1.2)	37.6 (1.4)	0.233	37.4–42.2	34.9–40.3	1,112	1,001
11–15 years old	68.2 (1.1)	55.8 (1.3)	0.000	66.0–70.4	53.3–58.3	780	702
16–17 years old	87.1 (2.3)	80.6 (2.4)	0.054	82.6–91.6	75.8–85.4	530	477

Source: List source and other notes here, including information on how sampling weights were applied to generate estimates.
Standard errors are shown in parentheses below the baseline and follow-up prevalence rates.

In presenting estimates of child labor in tables, verify that you have calculated prevalence rates using the correct formula and the correct denominator for your definition of child labor.

C. Presenting key results in clear tables and figures

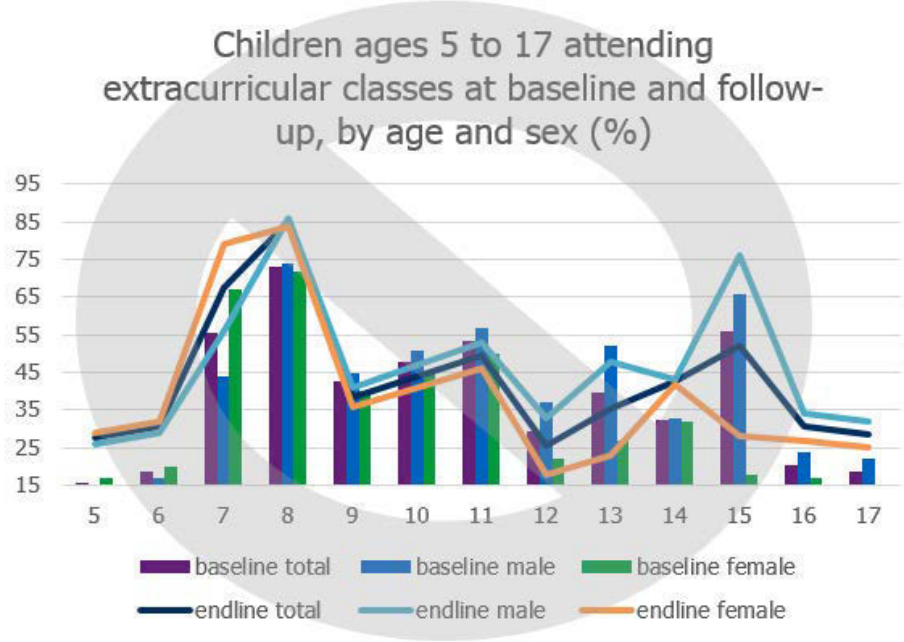
Make writing, tables, and figures as clear and concise as possible. Researchers may find the following rules of thumb helpful for enhancing the clarity of tables and figures in reports.

- **Make tables and figures easy to read**. Be sure that tables and figures are big enough and have sufficiently high resolution to be read easily. Avoid directly pasting figures from other documents, as this may result in blurred images. Use font sizes of 8 points or larger.
- **Explain features of tables and figures in a note below them.** For example, if some cells are blank or have dashes and no numbers, indicate if the blank cells or dashes represent a zero value, missing data, or data that do not apply. Clarify whether brackets represent confidence intervals. List data sources.
- **Ensure that percentages sum to 100 percent when appropriate.** When presenting categorical results, such as age ranges in which each person should fall into one category, make

sure that results sum to 100 percent. For example, when showing the age distribution of a sample of children ages 5–17, the percentage of children ages 5–10, 11–15, and 16–17 should add up to 100 percent. If results do not sum to 100 percent because of rounding, include a table note explaining this. Additionally, be sure subtotals add up to totals; if they do not, recheck the data. In some cases, individuals may fall into more than one category. For example, a child may select more than one reason for not attending school (such as school fees are too high and the school is too far). In such cases, percentages would not be expected to sum to 100 percent. When appropriate, use table notes to explain why subtotals do not sum to totals.

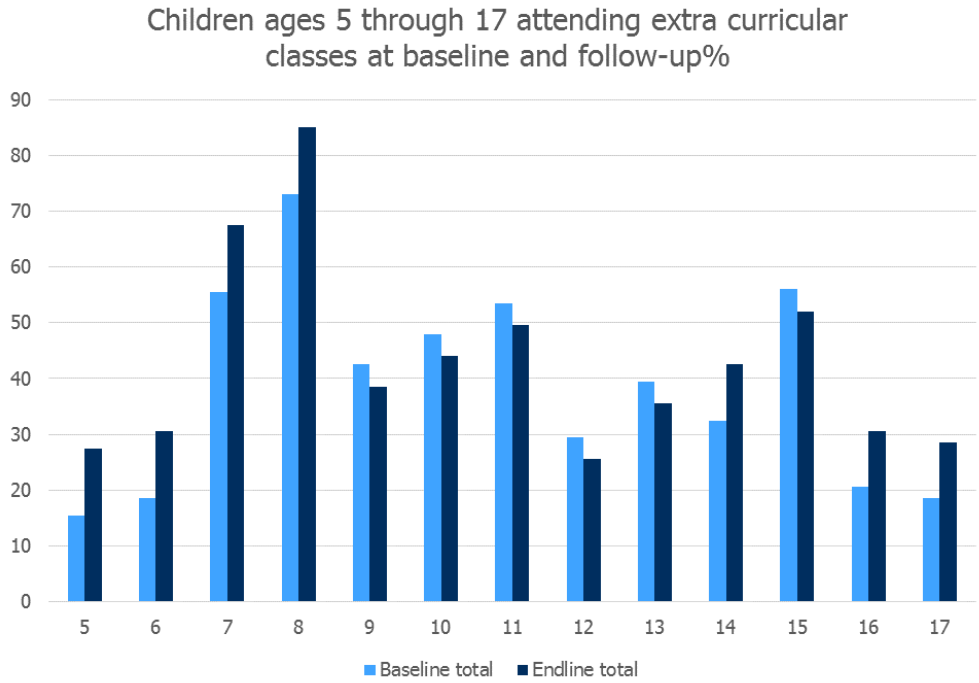
- **In general, avoid presenting numbers with more than one decimal place.** For means, proportions (including prevalence rates), standard errors, and confidence intervals, report one decimal place (for example, 5.1 percent). For counts, such as sample sizes, show whole numbers (for example, 430 people). For test statistics, such as *p*-values, report three decimal places.
- **When survey questions have only two possible answers, consider presenting results for only one category.** For example, if a survey question response is yes/no, reporting on the percentage who answered "yes" is sufficient. Omitting the "no" responses reduces the amount of information in a table, making it easier to read.
- **Number tables and figures sequentially.** Number tables and figures in the order in which they appear in the report.
- **Avoid overloading figures with too much information; use multiple figures if necessary.** An example of figure overload is below. Figure V.1 compares enrollment data of children at baseline and follow-up by gender and age. The figure is difficult to read and does not make any one point clearly; avoid using this approach. In contrast, it is easier to follow the main point of Figure V.2 because it is less cluttered. Figure V.2 compares the percentage of children attending extracurricular classes at baseline and follow-up at different age levels. Use your figures to convey information clearly. In general, convey one to two key points in a figure (as in Figure V.2); if figures include too much information (as in Figure V.1), it may be difficult for readers to identify the main point.

Figure V.1. Example that is overloaded with information (Do not follow this example)



Source: Developed as an example based on invented data.

Figure V.2. Example that presents the right amount of information (Follow this example)



Source: Developed as an example based on invented data.

- **Start the y-axis of figures at zero.** Start the y-axis at zero whenever it is appropriate for your data, to avoid visually distorting information. In Figure V.1 above, approximately 15 percent of 5-year-old children participated in extracurricular activities, although at first glance, it might appear that none did. Figure V.2 shows this more clearly.
- **Use bar graphs to compare statistics across categories, and line graphs to compare statistics across time.** Figure V.1 uses a line graph to show participation rates for children of different ages; however, most people might be inclined to interpret line graphs as representing changes over time. Figure V.2 avoids this confusion.
- **Report income in U.S. dollars.** Readers from different countries may not be familiar with the value of a local currency. Presenting monetary values in U.S. dollars and including the exchange rate for local currency makes the report accessible to an international audience.

To demonstrate a few of these rules of thumb, grantees may refer again to the templates for presenting child labor prevalence rates above (Tables 1 and 2). In these tables:

- ☑ Font sizes are larger than 8 points
- ☑ Table notes explain data source
- ☑ Subtotals of sample sizes sum to totals
- ☑ Percentages include just one decimal place, *p*-values have three decimal places, and counts are whole numbers

D. Statistical analysis

The statistical analysis described here is referred to as Step 4 in Chapter II on sampling. One of the first steps in conducting your analysis is to understand your data; report your findings, too. Report the number of households sampled, the number of those who were surveyed, and the number of children with complete child labor data. Additionally, calculate the household-level response rate (the percentage of households surveyed out of the number sampled) and child-level response rate (the percentage of complete child labor data out of all children in households sampled). Computations are described in more detail below:

Sampled households: The number of households sampled $\left(n_{households}\right)$.

If known, the number of children ages 5 through 17 within them (*n*).

Surveyed households: The number of the sampled households who were found by enumerators and consented to be surveyed $\left(n_{households\ surveyed}\right)$. The number of children ages 5 through 17 within the surveyed households $\left(n_{surveyed}\right)$.

The household-level response rate: $n_{households\ surveyed} / n_{households} * 100$.

Complete child labor data: The number of children ages 5 through 17 within the surveyed households for whom you have enough information to determine their child labor status $\left(n_{complete}\right)$.

The child-level response rate: $n_{complete} / n * 100$.

If the number of children in sampled households (including households that are not surveyed), n, is not known, estimate it as:

$$n_{households} * n_{surveyed} / n_{housheolds\ surveyed}.$$

The analysis of survey data (which we describe below) will not include children for whom you have incomplete data. For example, in calculating the sample proportion in child labor, children for whom you are missing information to determine their child labor status are removed from both the numerator and denominator.

If the completeness rate for child labor status is at least 80 percent, then no analysis adjustments are needed. If it is less than 80 percent, researchers can use nonresponse weights, possibly with help from a statistician (ILO Section 7.2.4).

Analysis—using data from the sample to estimate population quantities (ILO Sections 7.3.1 and 7.6.1). This text assumes equal probability sampling, which simplifies analysis, avoiding the need for survey weights. Consult the ILO manual for information about non-equal probability sampling and nonresponse weights.

When using equal probability designs, the estimate of an average or proportion in the population is just the average or proportion in the sample. To calculate the standard error for this estimate, the formulas take into account the similarity of children within PSUs. In cluster sampling, PSUs are the first groups of households that are sampled. In noncluster sampling, PSUs are the households. Calculation details are provided in the text box and implemented in the Excel template.

Calculation details

Let y_{hij} be the response for child *j* in PSU *i* (in stratum h). Let $y_{hi} = \sum_j y_{hij}$ be the total of the responses in PSU *i* (in stratum *h*). Let n_{hi} be the sample size in PSU *i* (in stratum *h*) and let *n* be the total sample size across all PSUs (and strata). Let n_{hI} be the number of PSUs sampled (in stratum h.

The estimate of the average or proportion is just the average or proportion in the data:

$$\bar{y} = \frac{\sum_h \sum_i y_{hi}}{n}$$

The standard error is calculated as:

$$se(\bar{y}) = \sqrt{\sum_h \sigma_h^2}$$

where

$$\sigma_h^2 = \frac{n_{hI}}{n_{hI} - 1} \sum_i (z_{hi} - \frac{z_h}{n_{hI}})^2$$

if $n_{h1} > 1$, and $\sigma_h^2 = 0$ otherwise. Define $z_{hi} = \frac{1}{n}(y_{hi} - \bar{y} n_{hi})$ and $z_h = \sum_i z_{hi}$.

For nonstratified sampling, drop the "*h*" and summation $\sum_h$ from the calculations.

Calculate a confidence interval as:

$$\bar{y} \pm 2 * se(\bar{y})$$

Estimating change over time

Researchers usually collect both baseline (before the intervention) and follow-up data. At baseline, the population of interest is all children ages 5 through 17 in a program's geographic target area before service provision starts. At follow-up, the population of interest is all children ages 5 through 17 in the same target area toward the end of the project. The population changes between baseline and follow-up as children age into and out of the target population and move into and out of the target area. Baseline and follow-up each have their own sampling frame because the population will have changed between baseline and follow-up data collection.

Use the same sampling design and survey tools at baseline and follow-up. If children's activities have a seasonable nature, for example, in agricultural settings, conduct data collection during the same season at both baseline and follow-up; otherwise, seasonal differences in child labor could distort before-and-after comparisons.

Confidence intervals for the change over time: Following the calculations above, obtain $\bar{y}^{baseline}$ and $se(y^{baseline})$ from baseline data and $\bar{y}^{follow-up}$ and $se(\bar{y}^{follow-up})$ from follow-up data. Estimate the

change over time as $\bar{y}^{follow-up} - \bar{y}^{baseline}$ and its standard error as $se(\bar{y}^{follow-up} - \bar{y}^{baseline}) = \sqrt{[se(\bar{y}^{follow-up})]^2 + [se(\bar{y}^{baseline})]^2}$ Calculate a confidence interval as:

$\bar{y}^{follow-up} - \bar{y}^{baseline} + 2 * se(\bar{y}^{follow-up} - \bar{y}^{baseline})$.

E. Comprehensive discussion of findings

Reports on child labor are more compelling if they include analysis that goes beyond stating the prevalence of child labor. Depending on the research questions being answered, some reports may include findings on key outcomes other than child labor, such as prevalence of the worst forms of child labor. Analysis may identify potential drivers of child labor and the implications for efforts to combat child labor in the context of the project. The following are strategies that researchers can follow in order to conduct such in-depth analysis of results.

- **Analyze relationships between child labor and other survey results.** Simple comparisons of child labor prevalence rates between subgroups using statistical *t*-tests might uncover important relationships between child characteristics and a propensity to participate in child labor. For example, it might be interesting to show how child labor participation varies by school attendance rates. If possible, researchers can also perform correlation analysis to identify variables that are most closely associated with child labor. Regression analysis may be used to show which household characteristics, such as wealth, household size, and child age, are most related to child labor. For example, one could regress an indicator of whether households had a child engaged in child labor on household characteristics to determine which variables have a large and statistically significant association with this outcome. However, when conducting such an analysis, keep in mind that characteristics that correlate with child labor do not necessarily cause child labor (that is, correlation does not imply causation). For example, it could be that eating fewer than three meals a day correlates with child labor. This correlation does not mean that eating fewer than three meals a day causes child labor. Rather, poverty is likely to be a contributing factor to both eating fewer than three meals a day and engaging in child labor.
- **Discuss surprising results and their implications for implementation.** Highlighting unexpected results provides an opportunity to reassess plans for implementation (if results are discovered at baseline) or to inform future work (if discovered at follow-up). For example, discovering that child laborers in project areas are more likely to go to school than child non-laborers at baseline could be a reason for implementers to focus program activities on supporting extracurricular activities rather than school enrollment.
- **Provide context.** Do not assume readers know the context of the project or country. It is helpful if the background section of the report briefly describes original plans for implementation, any changes to these plans, duration of activities, and project regions. This includes describing actual program implementation and events that caused disruptions that may have had large effects on the estimates of child labor, such as outbreaks of disease, natural

disasters, civil war, and so on. It is also sometimes helpful to describe cultural norms that are specific to the country so as to foster deeper understanding of the child labor prevalence.

- **Link qualitative results with survey results.** The findings from key informant interviews and focus groups are most useful when discussed in light of survey results. Analyzing qualitative findings in conjunction with survey results grounds the anecdotes in survey results, limiting analysts' ability to present selective anecdotes that simply demonstrate the success of the project. Good qualitative analyses may validate quantitative findings or may show patterns that are not obvious in the survey results. Similarly, survey results are most meaningful when interpreted in light of qualitative findings. Researchers may find it useful to discuss the consistency of findings from the quantitative survey and the qualitative interviews. For example, when discussing survey results on the relationship between schooling and child labor, you might use text boxes containing quotes from focus group discussions and key informant interviews in your report. Such quotes could provide further validation to the survey results and suggest reasons for the relationship between schooling and child labor. Qualitative results can also help researchers interpret unexpected survey results. For example, it would be surprising to find that children enrolled in school are more likely to also work. Focus groups could reveal that recruiters for child workers target children at their schools.
- **Indicate how results could influence implementation.** The data collection and analysis conducted to produce baseline and follow-up reports provide an opportunity for researchers and implementers to reflect on program implementation. At baseline, researchers typically discuss whether and how baseline findings support modifications to implementation plans. When doing so, use evidence in the data to back up any recommendations offered. At follow-up, discuss any implications findings have for future work to combat child labor. If a contractor conducts data collection and analysis on behalf of an organization implementing work to combat child labor, implementers should work closely with those contractors as they conduct their analysis to help them interpret data. If a separate organization is evaluating the effectiveness of the implementer, it would be appropriate to ensure that organization's independence and objectivity.

Exhibit V.1. Checklist for clear reporting

Report structure and content

1. The report includes the following key sections (see text for complete list):

- ☐ Front sections
- ☐ Background
- ☐ Data collection and analysis
- ☐ Child labor estimates
- ☐ Conclusions and recommendations

2. Key results are in the report and less important results are in an appendix or omitted.

- ☐ Includes context for specific intervention, as applicable (for example, education indicators for education projects), in main report
- ☐ Includes results on child labor indicators (child labor, hazardous child labor, children at risk of child labor, and worst forms of child labor [if applicable]) in main report
- ☐ Omits or places in an appendix less important contextual results

3. The report includes discussion of findings.

- ☐ Analyzes relationships between child labor and other survey results
- ☐ Discusses surprising results and their implications for project implementation, if applicable
- ☐ Provides context
- ☐ Links qualitative results with survey results

Presentation of child labor prevalence rates and other indicators

4. The report includes required statistics for prevalence rates.

- ☐ Includes estimated prevalence rates for child labor, children at risk of child labor, hazardous labor, and worst forms of child labor (if applicable)
- ☐ Includes standard errors
- ☐ Includes confidence intervals
- ☐ Includes sample sizes
- ☐ Describes how weights were used in generating estimates
- ☐ Follow-ups include the p-values from t-tests of the difference between baseline and follow-up prevalence rates

5. The report includes nonresponse rates.

- ☐ Includes nonresponse rates for key indicators, including variables used to identify child labor indicators
- ☐ Includes a description of nonresponse weights or other weights used in analysis

Tables and figures

6. Table and figure numbers are sequential and according to the order in which the tables and figures appear in the report.

- ☐ Numbering of tables and figures is sequential
- ☐ The list of tables and figures at the beginning of the report is accurate
- ☐ Text correctly and sequentially cites tables and figures

7. Tables and figures are easy to read.

- ☐ Font sizes are large enough to easily read (10-point font is best, but text, including table notes, should never be smaller than 8-point font)
- ☐ Images have sufficiently high resolution for printing
- ☐ Tables or figures inserted as pasted images are clearly legible

8. Percentages are accurate and consistent with other percentages reported in the same table.

- ☐ Percentages of categorical variables sum to 100 percent, when applicable
- ☐ Subtotals sum to totals

9. Table and figure notes and legends explain their features and sources.

- ☐ Tables and figures include data sources
- ☐ Table and figure titles or notes specify what sample was used for the analysis presented
- ☐ Tables and figures include labels for x- and y-axes and note the corresponding units used
- ☐ When appropriate, table notes explain what blanks or dashes in table cells represent (zero values, missing data, data that did not apply, or other)
- ☐ Table or figure notes or legends indicate what symbols or color coding mean—for example, if brackets in figures describe confidence intervals
- ☐ Table notes explain if subtotals do not add up to totals because of rounding

10. Tables and figures are concise, not overloaded.

- ☐ Complex tables or figures are separated into multiple tables or figures
- ☐ Figures do not use multiple y-axes
- ☐ For survey questions with only two possible answers, tables show results for only one response (that is, the percentage who answered "yes" but not also those who answered "no")

11. Figure types are appropriate for the type of data they represent.

- ☐ Bar graphs present data across categories (for example, by gender or region)
- ☐ Line graphs present data across time (for example, enrollment rates by year)

12. Figures do not visually distort information.

- ☐ Unless there is a reason not to, y-axes start at zero

13. Monetary values are reported in a currency that international readers understand.

- ☐ Equivalent values of local currency are reported in U.S. dollars
- ☐ A table or figure note includes the exchange rate used to convert local currency into U.S. dollars

14. Indicators use appropriate decimal places.

- ☐ Counts such as sample sizes use whole numbers
- ☐ Means, proportions (including prevalence rates), standard errors, and confidence intervals use one decimal place
- ☐ Test statistics, such as p-values, use three decimal places

VI. PREPARING PUBLIC USE DATA FILES

After collecting and analyzing data on child labor, many researchers choose to make their data available to the public online with no restrictions. This document describes the process for creating such public use files, which allow others to work with, and continue to learn from, the data. Because anyone can access public use files and users are not tracked, researchers must use caution to minimize the risk of disclosing respondents' identities either directly or indirectly. In contrast, restricted-use files are made available only upon request; those seeking access may be required to provide their name and institutional affiliation, describe the purpose of their research, and agree to terms of use. Researchers may allow more disclosure risk in restricted-use files.

Making data public enables other researchers to use the information. Researchers with access to public data can add to the body of knowledge relating to child labor and other issues. Moreover, they can compare their findings to those from other countries to learn about broader trends.

Creating public use files could, however, put survey respondents at risk. Making data public increases the risk that respondents' identities and private information could be disclosed and potentially misused. For example, if researchers publish health data in a way that makes survey respondents identifiable, it may be possible for some people to use the information to some respondents' disadvantage, such as to discriminate—in employment or housing decisions—against people with certain health histories.

Researchers must take care to protect respondents' identities when creating public use files. It is crucial to strike a balance between making data usable and preserving the privacy of respondents. In this document, we discuss some strategies for striking that balance. When in doubt, err on the side of protecting respondents, even if doing so reduces the usefulness of the public use data set. A data set that is made available to the general public must protect survey respondents even if, in doing so, the data suffer some loss of detail. A separate restricted-use file that retains more detail can be made available to researchers who request it and agree to follow privacy protocols.

In this document, we discuss strategies to ensure that public use data are useful to researchers and that safeguards are in place to protect individual respondents. We lay out a step-by-step approach to help researchers preserve the confidentiality of respondents' identities and information:

1. Understand data privacy issues
2. Develop a protocol for minimizing disclosure risk
3. Identify and mitigate risk of disclosure through direct identification
4. Identify risk of disclosure through indirect identification
5. Mitigate risk of disclosure through indirect identification
6. Consider advanced strategies for mitigating indirect disclosure risk
7. Document measures taken to minimize disclosure risk

We also provide a checklist that summarizes key steps.

Step 1: Understand data privacy

Preparing data for public use is not as simple as removing respondents' names and contact information. People preparing public use data files must protect their respondents not only from having their identities revealed directly, but also from having their identities and, therefore, their survey responses, revealed indirectly. For example, ill-meaning users could discover respondents' identities by combining geographic information with unique characteristics. An especially large family, a family with triplets, or a family with unusual livestock holdings would be easy to identify—even without names—if this information is available in combination with information about location. Data privacy measures will prevent this. In this section (Step 1), we introduce data privacy and related concepts.

To assess and mitigate the risk of disclosing respondents' personal or sensitive information, be sure your data collection team understands the following key concepts that relate to public use data:

- **Personally identifiable information (PII):** Any information that someone could use to identify a specific individual, such as name, address, or phone number.
- **Data confidentiality:** Protection from having a respondent's identity or potentially harmful information revealed. Examples include financial, work status, and health information.
- **Data disclosure:** The identification of an individual, either directly from the data or indirectly through connecting the data to other data sources. There are three types of data disclosure:
  - **Identity disclosure** is the release of data that explicitly identify an individual. An example would be the publishing of a list of participants in a job training program along with their names and addresses.
  - **Attribute disclosure** is the release of sensitive information attributable to a specific individual. An example would be the publishing of people's wages or other information about their work history.
  - **Indirect disclosure** is the release of information that potentially allows the identification of a specific individual. For example, if a respondent's profession and area of residence are uncommon, it might be easy to determine that person's identity.

Different types of protections may be required in different circumstances. Learn about and adhere to any relevant government regulations or legislation that relate to data confidentiality, security, and/or disclosure in your country. Each country has its own laws to protect research subjects, and researchers must follow them through all stages of the research process. Furthermore, if your data collection was reviewed by an Institutional Review Board (IRB) or similar body that considers the ethical implications of data collection and research, be sure that your public use files are consistent with the requirements set forth by the IRB.

Step 2: Develop a disclosure risk protocol

Develop a disclosure risk protocol to identify all points in the data collection process when confidentiality could be compromised. Because you might employ strategies to protect respondents' privacy early in the process, draft your disclosure risk protocol early. Key steps include the following:

- Review the survey instruments to identify potentially identifiable variables that are not necessary and could be eliminated. Eliminating unnecessary data from the survey and data will save work later. For example, if you do not need specific data on livestock owned, it may be best to eliminate such survey questions. In the disclosure risk protocol, call for eliminating unnecessary survey questions that would generate data that could be used to identify respondents.
- Identify points in the data collection process when respondents' identities could be linked to their survey responses.
    - Responses could be revealed if enumerators conduct interviews in public places or in parts of a house where respondents could be overheard by family members. In the disclosure risk protocol, call for conducting interviews in places where respondents will feel at ease to answer freely.
    - Responses could also be revealed if enumerators or interpreters are known to respondents. This is more likely to occur if enumerators are hired locally. In the disclosure risk protocol, provide strategies to ensure that enumerators are not known to respondents.
- Review external data sources to identify any that could be linked to your data set to identify respondents. For example, a student ID number could be used to identify a student's academic records. The names of villages could be used to link to publicly available data about a particular village. In the disclosure risk protocol, call for anonymizing such IDs.
- Identify variables most likely to be known by others. For example, neighbors or other community members are likely to be able to identify households with unusually large amounts of or unusual livestock. Similarly, they are likely to know households with unusual characteristics, such as triplets. In the disclosure risk protocol, include a process for identifying such variables and mitigating risk associated with them.

Throughout the process of assessing and mitigating risk, researchers must weigh the need to protect privacy against the utility of the data set to answer research questions. Some variables, such as phone numbers, are easily linked to individuals and add little or nothing to the utility of a data set for future research purposes. Such variables are easily removable from a public data set. Other variables, such as age and work history, are likely important for analysis; researchers must take care to keep as much of this useful information as possible without allowing identification of respondents. The steps in this document will help researchers assess these more difficult cases so they can decide how much information it is possible to preserve while still protecting respondents.

Step 3: Identify and mitigate risk of disclosure through direct identification

Some information in a data set might directly identify individual respondents. Examples of this type of information include names, addresses, identification numbers, and phone numbers.

Information that could directly identify respondents is easy for anyone to locate by reviewing the survey instrument or data set. The simplest way to avoid direct identification of subjects is to remove these variables from the data set and replace household or respondent identifiers with anonymous codes. If you need to maintain PII—for example, if you have to locate the same households or individuals for follow-up surveys—remove the PII from the public use file, assign anonymous IDs, and create a linking file that contains the PII and anonymous ID variables. Creating an anonymous ID variable ensures that the ID itself will not be correlated with or reveal anything about respondents. The linking file allows the researcher to link data from one data collection effort (such as baseline) to a future data collection effort (such as follow-up). Keep the linking file in a secure location separate from the rest of the data. Table VI.1 shows an example of how to use a linking file to keep PII separate from the public use data file.

An easy way to create the anonymous case ID is to sort your respondents in random order and create a new ID corresponding to the respondent's place in the list. To do this in STATA, follow these steps:

1. Sort your data in a way you can replicate, such as by the existing respondent ID.
2. "Set the seed." This allows you to replicate the random number generation command and generate the same set of random numbers. To do this, type `set seed 12345`, where 12345 could be any number of your choice.
3. Use the code `gen random=runiform()`, where "random" is the name of the random variable (you may use any name).
4. Sort your data by the random variable you created by typing `sort random`.
5. Create a new unique ID called "ID" by typing `gen ID=_n`.

Table VI.1. Linking files provide a link between PII and public use files

Full data set				Analysis file (for public use)		Linking file (for future surveys)		
Fname	Lname	Income	Linking case ID	Income	Linking case ID	First name	Last name	Linking Case ID
Jane	Doe	$500	12345	$500	12345	Jane	Doe	12345

Step 4: Identify risk of disclosure through indirect identification

Assessing the risk of indirect identification is less straightforward than simply removing sources of direct identification. Indirect identification can occur if data elements or combinations of data elements with few observations could be used to identify respondents indirectly or if the data set can be merged with outside information to discover attributes about individuals. In this section (Step 4), we present some ways indirect identification of individuals is possible. The next section (Step 5) describes potential strategies for mitigating disclosure risk.

- **Geographic information** could be used to identify respondents. Detailed information such as a specific village, city, or postal code increases the risk of identifying people in the data set, especially when combined with other attributes.
- **Information about the data collection process,** such as date of interview, interviewer name, interview location, or collection mode, could identify respondents. For example, community members might observe the date and time of an interview, which enumerators interviewed which respondents, or the method of data collection (for example, a tablet versus a paper-based survey). If the data set identifies which enumerator conducted the interview along with the date and mode, others might be able to identify the respondent.
- **External, publicly available data** could be combined with survey data to identify respondents. For example, if the respondent gives details about his or her school or employer, those details could be cross-referenced with public information about schools and businesses to identify the school or employer as well as the respondent's location.
- **Extreme values** of any continuous variable could be well known and could be used to identify respondents in an area. For example, in areas where cattle are raised, it may be common knowledge which households have the most cattle. Similarly, very old people in an area may be known.
- **Rare values or combinations of characteristics** (also known as "small cell sizes") could be used to identify people, particularly if these variables reflect characteristics that others are likely to know. Characteristics that could be used alone or in combination to identify people include, but are not limited to, age, race, gender, occupation, assets, and geographic location. For example, if date of birth and neighborhood are included, individuals could be identified by neighbors with knowledge of others' birthdays.

  To assess the risk of identifying people based on combinations of rare characteristics, tabulate any variables that could be used to identify people, such as date of birth, race, or occupation. A

general rule is to highlight any with fewer than three individuals. For example, detailed employment information may include some rare occupations. If there are only one or two priests in the survey area, labeling someone in the data as a priest would make it easy to identify him or her.

To assess whether combinations of variables can identify people, tabulate combinations of variables that could identify people and look for any cells with fewer than three individuals. For example, there may be 50 16-year-olds in the data set and 25 people who work in carpentry. Both are large enough numbers that there is no concern about inadvertent identification. But if the data set allows tabulations of age and type of work, there might be only two 16-year-olds who work in carpentry; the combination of the two variables could identify individual respondents. We demonstrate this in Table VI.2.

Table VI.2. Data elements with very few observations

	Age			
Occupation	**15**	**16**	**17**	**Total**
Agriculture	15	22	20	57
Carpentry	8	2	14	25
Commerce	20	25	25	70
Total	43	50	59	152

Step 5: Mitigate disclosure risk through indirect identification

After assessing possible disclosure risks in the data set, work to mitigate them. Here, we present some common methods for mitigating disclosure risk. Remember to document any changes you make to the data to mitigate disclosure risk.

- **Drop geographic identifiers**. To minimize the likelihood of indirect identification, consider dropping *all* geographic identifiers, including region, state, district, province, and village, as well as any other variables that might reveal respondents' location. Names of specific entities like schools or employers could also be dropped, as they are geographic identifiers and can reveal location.
- **Drop or recode variables that reveal geographic location.** Remove other information that might reveal respondents' location. For example, some variables reveal location-specific occupations (such as mining or fishing), crops that are grown only in certain regions, language or ethnic group, and so on. In many cases, such variables can be recoded to avoid geographic identification. For example, if one region of a country is known for rice and another for corn, both can be recoded as "staple crops." The exact variables and strategies for recoding will depend on the context of the specific country.

- **Remove variables** that could lead to identification of respondents. In addition to dropping geographic identifiers, look at any variables identified previously as potentially risky, including identifiable characteristics, characteristics that are strongly correlated with geographic identifiers, information about the data collection process, or variables that can be linked to other data sources. Consider which of those variables are unlikely to be useful for analysis. Retain information such as weights, primary sampling units (PSUs), or strata identifiers are likely to be necessary to replicate the analysis. If PSU IDs are also used for the national census or other purposes, change them to randomly generated IDs so that PSUs cannot be used to identify respondents' locations. Names of data entry technicians or enumerator supervisors, on the other hand, are unlikely to be important for analysis; remove them.
- **Remove detailed responses.** Some survey questions allow respondents to provide their own answers. Such response options are usually marked "Other (specify)." Convert responses like these, in which the respondent describes his or her own unique response, to "other" without the respondent's unique, word-for-word description.
- **Recode data** to reduce the risk of disclosure. Removing detail from some variables makes identifying individuals much more difficult. Below, we provide several methods for recoding data to reduce disclosure risk. Top-coding and bottom-coding, aggregating, and rounding are methods for reducing risk in continuous variables. Combining data into different categories is a method for categorical variables. Some of these methods change the data and will even slightly change mean values calculated with your data set. Although these changes are expected to be small, include them all in your documentation.
    - **Top-coding or bottom-coding extreme values of continuous variables** is a common strategy to remove detail on extreme values in the data, which could lead to respondents' identification. Consider top- and bottom-coding the top and bottom 5 percent of responses for all continuous variables that could be known by others. Age, assets (such as livestock or vehicles), house characteristics, and income are examples of continuous variables that could be known by others. Depending on the context, it may also be necessary to top- or bottom-code crop quantities and land assets.

      Table VI.3 provides an example of top- and bottom-coding a continuous variable. In this example, the lowest response option, zero, comprises 18 percent of observations. In this case, because the smallest category includes more than 5 percent of respondents, there is no need to recode the smallest values to the value that includes the 5th percentile. At the high end, the top 5 percent includes responses 17, 16, 14, 8, and 7—the 95th percentile falls in the response option of 7 cows (we know this because respondents with responses greater than 7 sum to less than 5 percent of respondents, and respondents with responses equal to or greater than 7 sum to more than 5 percent of respondents). Recode all response options above the 95th percentile (here, 8, 14, 16, and 17) to the value that includes the 95th percentile (in this case, 7). You can identify the response option that

includes the 95th percentile manually or use your statistical software to identify the value at the 95th percentile. In STATA, for a variable called "cows," use the code `sum cows, detail` to see the value that includes the 95th percentile. Table VI.3 provides an example of how a variable describing how many cows a respondent owns would be recoded.

Table VI.3. Top-coding continuous variables above the 95th percentile

[?] Before This version is *unlikely* to protect respondents from disclosure risk		
Number of cows	**Count**	**Percentage of total**
0	205	18
1	134	12
2	192	17
3	188	17
4	144	13
5	102	9
6	78	7
7	44	4
8	25	2
14	3	0
16	2	0
17	2	0

Note: Because of rounding, percentages do not sum to 100 percent.

[✓] After This version *does* protect respondents from disclosure risk		
Number of cows	**Count**	**Percentage of total**
0	205	18
1	134	12
2	192	17
3	188	17
4	144	13
5	102	9
6	78	7
7	76	6

Note: The revised table shows that all 32 observations with values above the 95th percentile have been recoded to the value at the 95th percentile, which is 7.

- **Aggregating data** into bands or percentiles is a simple way to remove some detail for continuous variables. For example, if reporting the number of rooms in respondents' homes produces some cells with very few observations—either alone or when combined with other data—aggregating rooms into bands, such as one to three rooms, four to six rooms, and so on, is an acceptable approach. We show an example of this approach in Table VI.4. A disadvantage of this approach is that it can convert a continuous variable into a categorical variable, and categorical variables are more difficult to analyze. Think through your options carefully to weigh the pros and cons of recoding a certain variable. Think about the cutoffs or categories relevant for analysis; for example, if recoding hours worked into categories, make sure that the hour categories match what is needed for determination of child labor status.

Table VI.4. Aggregating numeric variables with very few observations

[?] Before This version is *unlikely* to protect respondents from disclosure risk	
Rooms in house	**Count**
1	5
2	48
3	55
4	50
5	37
6	8
7	12
8	6
9	2

[✓] After This version *does* protect respondents from disclosure risk	
Rooms in house	**Count**
1–3	108
4–6	95
7–9	20

Note: Specific values in the original table on the left, which could be known by others, were combined into less specific ranges shown in the revised table on the right.

- **Rounding all values** is another method for removing detail from continuous variables. If, for example, exact wages might help to identify individuals, rounding to the nearest $1 is an appropriate approach. Individuals who earn $2.65 per day, for example, would be recoded to $3.
- **Combining data** by moving observations into other existing categories or creating new categories can remove detail from categorical variables.

  For ordinal variables, such as educational level, neighboring categories may be combined. For example, if only two respondents in the data set earned a postgraduate degree, combine them with those who earned college degrees and rename the new category "college degree or greater."

Table VI.5. Combining categories for variables with very few observations

[?] Before This version is *unlikely* to protect respondents from disclosure risk	
Educational attainment	**Count**
Less than primary	8
Primary	20
Secondary	43
High school	45
College	55
Graduate degree	2

[✓] After This version *does* protect respondents from disclosure risk	
Educational attainment	**Count**
Less than primary	8
Primary	20
Secondary	43
High school	45
College or greater	57

Note: The original table on the left shows that the graduate degree category has only two respondents. To mitigate the disclosure risk associated with having only two respondents in a category that might be known by others, the "graduate degree" and "college" categories have been combined into a new "college or greater" category in the revised table on the right.

Data elements with unranked categories can be combined into larger categories. For example, if there are only two hairdressers in the data set, combining similar professions into one category is reasonable. Geographic indicators can be aggregated up to a larger scale; for example, if the data identify small villages, grouping villages by district or state is acceptable.

If small categories are too dissimilar to combine into existing groups, recoding as "other" is a common approach. For example, if the data set asks why children miss school and only two respondents report missing school because of a family member's illness, there may not be another similar category relevant for the response. In this case, to avoid identifying those students, their responses could simply be recoded as "other."

As you assess different options, consider carefully how each approach affects the value of the data set for future research, especially in the context of child labor definitions and related topics. In some cases, you may be able to recode one variable instead of another. For example, in a case in which there were only three 8th-graders working in agriculture, you could recode either grade (into a categorical variable) or occupation (combining rare occupations) to avoid having only three individuals in that cell. In this example, recoding grade to school level may be the preferred option, so that detail on the occupation variable is not lost. On the other hand, if specific grades are particularly important to your research questions, you may choose instead to recode occupation. In general, age is a high-priority variable to keep for child labor research.

It can be difficult to strike a balance between protecting individuals' information and preserving the usefulness of your data set. Think carefully about your options so that you choose approaches to protecting information that will preserve the research value of your data set as much as possible. When in doubt, err on the side of protecting privacy. If detail is lacking, researchers may be able to request a more detailed version and sign an agreement to use it.

Step 6: Consider other advanced strategies for mitigating disclosure risk

The methods we discussed in the previous sections involve removing variables or reducing the level of detail in the data set. Below, we describe some advanced methods for reducing disclosure risk while minimizing the amount of information lost. Depending on the possible risks in your data and your staff's capacity to implement these methods, you may want to consider using them. Details on how to use these strategies are available online in the United States Federal Committee on Statistical Methodology's Statistical Policy Working Paper 22, "Report on Statistical Disclosure Limitation Methodology."

- **Data swapping** involves matching each observation to another observation in the data set, based on a set of primary respondent attributes. The values of other variables are then swapped between those matched observations.
- **Data shuffling** involves several steps. First, the data are rank ordered from lowest to highest on the values of the variable that will be "shuffled." This could be done for continuous variables, such as earnings, that could be used to identify respondents. The variable's values are then modified slightly using another method, such as random rounding. Next, the modified data are sorted from lowest to highest. Finally, the modified value is replaced with the original value that has the same rank as that of the modified data value.
- **Data blurring** is a method of grouping together small groups of respondents who are similar based on a set of chosen characteristics. This method involves calculating the aggregate value of a variable for that group and replacing the value of the variable for each individual with the aggregate value of the group.
- **Data substitution** involves substituting random values from a different file for values of a variable, such as substituting random male names for all males and female names for all females in the database to protect the original name but maintain the distribution and appearance of the data.

Step 7: Document all measures taken to minimize disclosure risk

Include full, clear documentation of any changes made to the data set, specifying all variables and combinations of variables that were checked for possible risks, as well as those that were determined to be a risk. List all variables they have been removed or edited (and describe any edits) and highlight revised variable values that arise from grouping categories together (indicating which previous

categories correspond to each new category). We provide details about documentation in the "Data Format and Validity" and "Analysis and Reporting" chapters.

Researchers may use program files showing how the original data file was edited. Program files can easily demonstrate the changes so that anyone using the data knows how the public data they are using differ from the original.

Exhibit VI.1. Checklist for public use data files

1. Develop disclosure risk protocol

- ☐ Protocol includes all variables that researchers will check for possible risk of direct identification. Common personally identifiable information (PII) to check for and drop include the following:
  - **i.** Names
  - **ii.** Addresses
  - **iii.** Phone numbers
  - **iv.** Email addresses
  - **v.** Identification numbers, such as national ID or passport numbers
  - **vi.** Birth dates
- ☐ Protocol includes the identification of other variables that could be used to identify individuals indirectly. At a minimum, the protocol calls for reviewing combinations with these variables that can be easily observed:
  - **i.** Age
  - **ii.** Sex
  - **iii.** Race or ethnicity
  - **iv.** Language
  - **v.** Occupation or workplace
  - **vi.** Educational status
  - **vii.** Crops grown
  - **viii.** Publicly known assets, such as land, livestock, vehicles, or farm equipment owned
  - **ix.** Variables that describe the interview process, such as interview date, time, location, method, or enumerator name
- ☐ Protocol includes a review of survey instruments to identify and remove unnecessary survey questions.
- ☐ Protocol includes strategies to minimize disclosure risks during data collection, including proper selection of interview locations and staffing to avoid using enumerators known to respondents.
- ☐ Protocol proposes methods for mitigating disclosure risks identified.
- ☐ Protocol is consistent with requirements of IRB, if applicable.
- ☐ Researcher understands any relevant national laws regarding disclosure risk, and protocol follows legal requirements.

2. Mitigate various types of disclosure risk

- ☐ Remove PII that directly identifies respondents.
- ☐ Create new, randomly generated household IDs and store them, along with names, in a linking file in a secure location. If the data set also includes PSUs used by others, such as a national statistics office, replace PSUs with randomly generated PSU identifiers.
- ☐ Remove all geographic indicators, including region, state/province, district, municipality, city, village, and neighborhood. If not removing geographic identifiers, conduct a thorough review of all other variables that might be used to identify respondents indirectly in combination with geographic identifiers.
- ☐ Remove variables that are closely correlated with geographic identifiers, such as ethnic group, language, crops grown, or geography-specific industries.
- ☐ Remove names of schools, employers, and other institutions that could be linked to a specific location.
- ☐ Extract details about the data collection process.
- ☐ Consider possible links to other publicly available data. If applicable, replace each PSU code with a randomly generated number.
- ☐ Recode all "other (specify)" responses to "other."
- ☐ For continuous variables that may be known to others, top-code values above the 95th percentile to the value that includes the 95th percentile and bottom-code values below the 5th percentile to the value that includes the 5th percentile. If the smallest (largest) values include 5 percent of responses or more, there is no need to recode the smallest (largest) values.
- ☐ Eliminate cells with fewer than three respondents, either for individual variables or combinations of variables that could identify individuals.

3. Check that all edits made to the data are documented clearly

- ☐ Include variables and combinations of variables that were checked for possible disclosure risk.
- ☐ Explain mitigation strategy used for each risk.
- ☐ Include in documentation the program files used to assess and mitigate disclosure risk. This program file may serve as the documentation of variables and combinations of variables that were checked and the mitigation strategy used for each risk.

VII. ONLINE RESOURCES, BY CHAPTER

This chapter provides links to online resources offering in-depth information on the topics touched on in these documents.

A. Sampling

Comprehensive resource on sampling design for household-based surveys on child labor:

International Labour Organization. *Sampling for Household-Based Surveys of Child Labour.* Geneva: International Labour Organization, 2008. Available at http://www.ilo.org/ipec/Informationresources/WCMS_099362/lang--en/index.htm. Accessed September 26, 2018.

- This manual provides a detailed description of sampling approaches for household-based surveys of child labor. The sampling text includes references to specific sections of the manual that may be of use to researchers designing a survey to estimate the prevalence of child labor.

Resource on sampling elusive populations:

International Labour Organization. Sampling *Elusive Populations: Applications to Studies of Child Labour.* Geneva: International Labour Organization, 2014. Available at http://www.ilo.org/ipec/Informationresources/WCMS_IPEC_PUB_25535/lang--en/index.htm. Accessed September 26, 2018.

- This guide is a complement to the ILO resource on sampling listed above and provides advice on how to draw a sample from a population when a complete sample frame is unavailable or some members of the population may be missing from the sample frame.

Interactive resources on sampling and analysis:

International Labour Organization. "ILO-IPEC Interactive Tools in Sampling with Household-Based Child Labour Surveys." Available at http://www.ilo.org/ipec/ChildlabourstatisticsSIMPOC/Manuals/WCMS_304559/lang--en/index.htm. Accessed September 26, 2018.

- These interactive tools could be useful for researchers using a sampling design that differs from the basic approach described in our sampling resources. Tool 7 provides an approach to estimating sampling errors that could be useful to researchers following the basic design.

B. Survey development and data collection

Comprehensive resources on survey development and data collection:

International Labour Organization. *Child Labour Statistics: Manual on Methodologies for Data Collection Through Surveys*. Geneva: International Labour Organization, 2004. Available at http://www.ilo.org/ipec/Informationresources/WCMS_IPEC_PUB_141/lang--en/index.htm. Accessed September 26, 2018.

- This manual describes methods for administering child labor surveys and analyzing their results, including survey design, enumerator training, sampling, data processing, and analysis.

Survey Research Center. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan, 2016. Available at http://www.ccsg.isr.umich.edu/index.php/chapters. Accessed September 26, 2018.

- The tenders, bids, and contracts chapter describes searching for, identifying, and hiring a data collection firm.
- The questionnaire design chapter describes questionnaire design, with a focus on developing appropriate content that will be well understood by respondents.
- The instrument technical design chapter describes developing a technically functional survey for paper-based or electronic data collection.
- The survey quality chapter includes a description of various techniques to improve survey quality.
- The interviewer recruitment, selection, and training chapter provides tips for finding and preparing high quality enumerators.

Grosh, Margaret, and Paul Glewwe (eds.). *Designing Household Survey Questionnaires.* Washington, DC: The World Bank, 2000. Available at https://openknowledge.worldbank.org/handle/10986/25338 (Volume 1), https://openknowledge.worldbank.org/handle/10986/15194 (Volume 2), and https://openknowledge.worldbank.org/handle/10986/15195 (Volume 3). Accessed September 26, 2018.

- This three-volume, detailed guide is old but includes information that is relevant today, based on 15 years of experience implementing the World Bank's Living Standards Measurement Surveys. Volume 1's most relevant chapters include metadata (data about how a survey was administered), household rosters, education, employment, and transfers and other non-labor income, among others; volume 2's most relevant chapters are on savings, credit, and time use. Volume 3 includes draft survey modules for each of the topics listed here, as well as others.

World Bank. "Living Standards Measurement Study Multimedia." Washington, DC: World Bank, 2018. Available at http://surveys.worldbank.org/lsms/multimedia. Accessed September 22, 2018.

- The World Bank provides online training modules on various aspects of household survey design, implementation, and analysis through videos and PowerPoint presentations.

Sample survey:

International Labour Organization. "SIMPOC questionnaires for stand-alone national child labour surveys" Geneva: International Labour Organization, 2007. Available at http://www.ilo.org/ipec/Informationresources/WCMS_IPEC_PUB_4946/lang--en/index.htm. Accessed September 26, 2018.

- ILO implements the SIMPOC survey around the world to collect data on child labor in various country contexts. The survey is available online and includes adult, household, and child modules. Researchers may consider using it as a starting point for developing their own survey.

Electronic data collection:

World Bank. "Seminar on Survey Solutions: An Innovative Tool for Surveys." Available at http://www.worldbank.org/en/events/2015/06/08/survey-solution. Accessed September 26, 2018.

- Survey Solutions is a free program available online for programming for electronic data collection. This webpage includes a link to video presentations on how to use the software and a link to access the web-based software.

United States Census Bureau. "Census and Survey Processing System (CSPro)." Available online at https://www.census.gov/data/software/cspro.html. Accessed September 26, 2018.

- CSPro is another free program available online for programming for electronic data collection. This webpage provides information on how to use the software as well as a link to download the software.

List randomization for asking sensitive questions:

Karlan, Dean, and Jonathan Zinman. "List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds." *Journal of Development Economics,* vol. 98, 2012, pp. 71–75. Available at http://deankarlan.com/wp-content/uploads/2018/01/listrandom.pdf. Accessed September 22, 2018.

- This journal article describes techniques to gather data on sensitive behaviors, including a technique known as list randomization. With this technique, all respondents are shown a list of statements and asked how many are true for them. A randomly selected subset of respondents sees a statement that other respondents do not see, such as "You engage in

illegal behavior." The analyst can estimate the percentage of the sample that engages in an illegal behavior by subtracting the average number of responses identified as true in the group that sees the additional statement from the average number of responses identified as true in the group that does not see the additional statement.

Data collection with children:

De Leeuw, Edith, Natacha Borgers, and Astrid Smits. "Pretesting Questionnaires for Children and Adolescents." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith Tessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken: John Wiley and Sons, Inc., 2004. Available at https://dspace.library.uu.nl/bitstream/handle/1874/44538/EdL-pretesting+2004.pdf?sequence=1. Accessed September 26, 2018.

- This chapter describes children's ability to understand and respond to survey questions at different ages and includes strategies for writing age-appropriate questions.

C. Data format and validity

Resource on data format and validity:

Survey Research Center. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan, 2016. Available at http://www.ccsg.isr.umich.edu/. Accessed September 26, 2018.

- The data processing and statistical adjustment chapter covers issues related to data cleaning and formatting. The first section in the chapter (Section 1) covers coding data. Points 1.4 and 1.5 cover how to develop a data dictionary and a codebook, respectively.

D. Analysis and reporting

Resources on child labor analysis and reporting:

International Labour Organization. *Manual for Child Labour Data Analysis and Statistical Reports.* Geneva: International Labour Organization, 2004. Available at http://www.ilo.org/ipec/Informationresources/WCMS_IPEC_PUB_3079/lang--en/index.htm. Accessed September 26, 2018.

- This manual includes advice on analysis and reporting of child labor statistics.

International Labour Organization. National Child Labour Survey Reports. http://www.ilo.org/ipec/ChildlabourstatisticsSIMPOC/Questionnairessurveysandreports/lang--en/index.htm. Accessed September 26, 2018.

- This webpage includes links to numerous national reports based on surveys of child labor that researchers may find useful to consult.

General resources on presenting clear tables and figures:

United Nations Economic Commission for Europe. *Making Data Meaningful, Part 2: A Guide to Presenting Statistics*. Geneva: United Nations, 2009. Available at http://ec.europa.eu/eurostat/documents/64157/4374310/33-UNECE-making-data-meaningful-Part2-EN.pdf/d5b954e2-b110-469b-a2b5-78aa7c12ab62. Accessed September 26, 2018.

- This 58-page guide presents tips on presenting statistics in ways that are easily understood and on preparing clear and effective tables and figures.

United Kingdom Governmental Statistics Office. *Effective Tables and Graphs in Official Statistics: Guidance for Producers*, 2nd edition. London: United Kingdom Governmental Statistics Office, 2017. Available at https://gss.civilservice.gov.uk/wp-content/uploads/2017/01/GUIDANCE-document-Effective-charts-and-tables-in-official-statistics-Version-2.0-Jan2017.pdf. Accessed September 26, 2018.

- This 36-page document provides examples of clear tables and graphics.

Evergreen, Stephanie. "Evergreen Data: Intentional Reporting and Data Visualization." Available at http://stephanieevergreen.com/. Accessed September 26, 2018.

- This website offers ideas for making reports and other documents visually appealing while conveying the author's key points. It offers several handouts, including a data visualization checklist and a light-hearted document with tips on making long reports easier to read. The author's blog contains entries on data visualization that may offer ideas for researchers' reports.