**Assessing Progress in Reducing Child Labor in Cocoa Production in Cocoa Growing Areas of Côte d’Ivoire and Ghana – 2018/2019 Survey Datasets Read-me**

**Prepared by NORC at the University of Chicago, for DOL’s ILAB**

This document describes the STATA datasets, STATA do files, R scripts, word documents, and excel workbooks prepared for the Final Report.

* If you would like to use the STATA datasets, you can simply download those and use them immediately without needing to run the do file or any R scripts.
* The STATA do file (Descriptive Tables.do) is the only file you need to run to replicate the point estimates found in the report.
* The R scripts (Create R datasets & R significance tests) only need to be run if you want to replicate the significance tests found in the report, but these files can take multiple days to run due to the bootstrap approach that is outlined below.

The “Datasets and Codes” sub-folder contains five STATA datasets and one STATA do file (Descriptive Tables.do) that creates the descriptive tables included in the report. The datasets include the cleaned survey data with the labelled final indicators used for the report, in STATA format. Personally identifiable information (PII) have been removed from the datasets to protect the confidentiality of respondents. All numbers in the report can be replicated by running the STATA do file (Descriptive Tables.do).

In addition, the “Datasets and Codes” sub-folder contains two R scripts that prepare the STATA data to be used in R and run the significance tests for the report tables. All numbers in the report have been replicated in STATA and R, and all significance tests have been run only in R. If you want to replicate each statistical significance test of difference between two survey rounds, you need to use the R codes (Create R datasets & R significance tests). Please note that the statistical significance tests use Rao-Wu bootstrap standards errors for generating the test statistics and takes considerable time to draw the Rao-Wu bootstrap standards errors for each individual test. Our estimates indicate that the time required to complete all the tests using the two R files would be approximately 2-4 days with a standard computing facility. The R script does not need to be run unless you want to replicate the significance tests from the report. Finally, the “Datasets and Codes” sub-folder contains an excel file with all the relevant information to replicate the 2018 weights, which is used in the R scripts to create the final R dataset.

The “Questionnaires and Crosswalk” sub-folder contains the surveys used in 2018 for the household roster, head of household, child, cocoa shed, school, and community leaders. A crosswalk was also added into the “Questionnaires and Crosswalk” sub-folder to show how the key child labor and hazardous child labor indicators were created from the questions in the survey.

Each STATA dataset file (in the “Datasets and Codes" sub-folder) is listed below along with notes on content:

1. all\_surveys\_2018\_final.dta
   * Includes information from the roster, head of household, child, school, community leader, and cocoa shed surveys.
   * The school, community leader, and cocoa shed surveys were reshaped, and indicators were constructed at the enumeration area level (i.e. each variable has one observation per enumeration area). The surveys were linked using the variable ‘EA\_2018’, which acts as a unique identifier for the enumeration areas. Note that some enumeration areas have more than one survey while others do not have any, depending on availability of interviewees.
   * The roster, head of household, and child surveys were linked using a variable that allowed us to identify the households in each enumeration area. There is one roster and household head survey per child survey, but are often multiple child surveys for each roster and household head (one for each child in the household).
   * The variable ‘weight’ has the survey weights to be used for any estimations at the child level for agricultural households. The variable ‘cocoa\_weight’ has the survey weights to be used for any estimations at the child level for cocoa households.
2. child\_survey\_2008\_2013\_2018\_final.dta
   * Includes the appended datasets of the child surveys conducted in 2008/09, 2013/14, and 2018/19. Note that the surveys have changed slightly across survey rounds and not all the same variables are available for each round.
   * The 2008/09 and 2013/14 rounds were collected by Tulane, while NORC led the data collection efforts of the 2018/19 round.
   * The variable ‘Year’ includes the year in which the data was gathered.
   * The variable ‘weight’ has the survey weights to be used for any estimations at the child level for agricultural households. The variable ‘cocoa\_weight’ has the survey weights to be used for any estimations at the child level for cocoa households.
3. HHH\_all\_years\_final.dta
   * Includes the appended datasets of the household head surveys conducted in 2008/09, 2013/14, and 2018/19. Note that this only includes the relevant variables for creation of the head of household tables and not all survey questions.
   * The 2008/09 and 2013/14 rounds were collected by Tulane, while NORC led the data collection efforts of the 2018/19 round.
   * The variable ‘Year’ includes the year in which the data was gathered.
   * The variable ‘weight’ has the survey weights to be used for any estimations at the child level for agricultural households. The variable ‘cocoa\_weight’ has the survey weights to be used for any estimations at the child level for cocoa households.
4. Roster\_2008\_2018\_final.dta
   * Includes the appended datasets of the roster surveys conducted in 2008/09 and 2018/19. Note that this only includes the relevant variables for creation of the roster tables and not all survey questions.
   * The 2008/09 round was collected by Tulane, while NORC led the data collection efforts of the 2018/19 round.
   * The variable ‘Year’ includes the year in which the data was gathered.
   * The variable ‘weight’ has the survey weights to be used for any estimations at the child level for agricultural households. The variable ‘cocoa\_weight’ has the survey weights to be used for any estimations at the child level for cocoa households.
5. Sample\_2018\_final.dta
   * Includes the appended datasets of the listing conducted in 2018/19. Note that this only includes the relevant variables for creation of the sample tables and not all survey questions.
   * NORC led the data collection efforts of the 2018/19 round.

STATA dataset notes:

* The data come from a combination of six different surveys that are all fairly long and complex with skip patterns. We suggest that you thoroughly review the surveys when anything in the data is not as you expect to make sure you fully understand what each variable corresponds to in the survey and what skip patterns may be important to consider.
* Most variables created after data collection have the term “7d” or “12m” as part of their names. These terms refer to the 7 days and 12 months periods, respectively. When the name does not include either of these terms, the variable was constructed for the 12 month reference period.
* In the variable labels, the following terms mean “child engaged in cocoa production”: “cocoa child”, “cocoa children”, “child in cocoa”, “children in cocoa work”, and “child working in cocoa”.
* In the variable labels, the following terms mean “child working in agriculture”: “child in agriculture”, “child engaged in agriculture”, and “child involved in agriculture”.
* In the variable labels, “child in non–cocoa agriculture” means “child working in non–cocoa agriculture”.
* For the variables on cocoa hours, this is the number of hours engaged in any economic activity for children who work in cocoa production. These are not measures of the number of hours spent in cocoa production.

The STATA do file (in the “Datasets and Codes” sub-folder) to create the report tables:

1. Descriptive Tables.do
   * Please download all the STATA files (5 datasets and 1 do file) to the same folder, and replace the directory in line 6 of the do file with the route of this folder.
   * When you run the do file, a series of estimates will be exported to the file called "Shell Tables.xlsx" in the same folder.
   * You cannot run the do-file (i.e. you will get an error message) if you have the excel file open.
   * This file will take 1-2 hours to run depending on your computer.

Each word documents (in the “Questionnaires and Crosswalk” sub-folder) is listed below along with notes about their contents:

1. Child Questionnaire
   * This document contains all questions in English asked to children for the child questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” and “all\_surveys\_2018\_final.dta” STATA files, except for any variables that contain PII.
2. Cocoa Shed Questionnaire
   * This document contains all questions in English asked to cocoa shed workers for the cocoa shed questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file, except for any variables that contain PII.
3. Community Leader Questionnaire
   * This document contains all questions in English asked to community leaders for the community leader questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file, except for any variables that contain PII.
4. Head of Household Questionnaire
   * This document contains all questions in English asked to household heads for the household head questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file, except for any variables that contain PII. The variables used to create the household head tables can also be found in “HHH\_all years\_final.dta”.
5. Household Roster Questionnaire
   * This document contains all questions in English asked to household heads for the household roster questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file, except for any variables that contain PII. The variables used to the roster tables can also be found in “Roster\_2008\_2018.dta”.
6. School Questionnaire
   * This document contains all questions in English asked to school leaders for the school questionnaire in 2018 and includes skip patterns.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file, except for any variables that contain PII.
7. Child Labor CrossWalk Ghana\_CDI\_Common
   * This document contains a crosswalk with the child labor definitions for Ghana, Cote d’Ivoire, and the common definition.
   * The variable names can be mapped to the “child\_survey\_2008\_2013\_2018\_final.dta” STATA file and the “Child Questionnaire” word document.

Each excel file (in the “Datasets and Codes” sub-folder) is listed below along with notes about their contents:

1. Shell Tables
   * This excel contains all the descriptive tables in the report, and the point estimates are automatically populated by running the “Descriptive Tables.do” STATA file.
   * The significance tests from this excel can be replicated in the “R significance tests.R” R script.
2. Ghana-CDI\_Survey\_Weight\_Count
   * This excel contains the data used for the creating the sample weights in 2018.
   * This file is used by the R script “Create R datasets.R” to replicate the weights used in STATA.

The R scripts (in the “Datasets and Codes” sub-folder) used to conduct the significance tests are listed below:

1. Create R datasets.R
   * Please download the “Ghana-CDI\_Survey\_Weight\_Count” excel to the same folder with the STATA datasets, and replace the directory in line 13 of the R file with the route of this folder.
   * When you run the R file it will create the R dataset “Update Tables.Rdata” in the same folder that is ready to be used in the “R significance tests.R” file.
   * This file will take around 1 hour to run depending on your computer.
2. R significance tests.R
   * Please download this to the same folder with the STATA files and R dataset, and replace the directory in line 11 of the R file with the route of this folder.
   * When you run the R script it will run the significance tests for each table in the “Shell Tables” excel. The significance levels will not be automatically populated into excel.
   * This file will take multiple days to run depending on your computer.