



## **STEP Household Questionnaire**

### **Guidelines for Data Processing**

**This Version: December 11, 2012**

Table of Contents

1. Data Entry Process and Timing..... 3

2. Data Files Structure ..... 4

3. Consistency Checks ..... 5

4. Appendix..... 15

The present document provides guidelines for processing data collected through the STEP Household Questionnaire.

It is specifically intended for Project Managers and all Data Entry Staff, regardless of the Data Entry Program (DEP) they are using, and details:

- (i) the data entry process and timing;
- (ii) the structure data files must comply with before being submitted to the World Bank STEP Core Team;
- (iii) data checks all Survey Firms are expected to carry out during fieldwork and data cleaning.

The guidelines are organized in three points:

- (i) **Standards**, which describe particular norms the Survey Firm is expected to comply with;
- (ii) Information to provide *During Fieldwork*;
- (iii) Information to provide in the *Final Implementation Report*.

Survey Firms are asked to provide either the **full list** or the **frequency** of particular inconsistencies:

- A full list of inconsistencies is required in cases that are unlikely to occur;
- Whereas only frequencies are required in cases that are more likely to present errors.

## 1. Data Entry Process and Timing

**Standard 1.** STEP Technical Standards allow Survey Firms implementing the STEP Survey to develop their own Data Entry Program (DEP) for the Household Questionnaire, if approved by the STEP Consortium.

Survey Firms developing their own DEP should use the present document as a guideline detailing specifications their DEP should comply with.

**Standard 2.** If the Survey Firm develops its own DEP, it must inform the STEP Consortium of the platform type it intends to use in the National Survey Design Planning Report (NSDPR).

**Standard 3.** The DEP will include a double-data entry process.

Two different operators will enter the data from the Household Questionnaire. All differences in the captured data will be resolved by the second data entry operator, as the DEP will flag differences with the data entered by the first data entry operator.

If the Survey Firm develops its own DEP and uses another method to reconcile differences, it will provide the STEP Consortium with a detailed description of its reconciliation process.

**Standard 4.** If a Survey Firm chooses to develop its own DEP, once it is ready and before the pilot, the Survey Firm should provide the STEP Consortium with made-up data used to test the DEP and check the structure and format of the data.

The data files should be sent in STATA or SPSS.

**Standard 5.** Within a week of completing the Pilot Survey, the Survey Firm will provide the STEP Consortium with data files based on data collected during the Pilot.

These data files should respect all standards required by the STEP Consortium.

**Standard 6.** Data entry will start no longer than 5 days after the start of fieldwork, to provide feedback to the field and enable Supervisors and/or Interviewers to revisit respondents whenever required.

**Standard 7.** The Survey Firm is expected to transmit the first week of data to the STEP Consortium within two weeks of the start of the fieldwork and the second week of data within the first three weeks of fieldwork.

**Standard 8.** If raw data cannot be delivered under the STEP format and structure during fieldwork or if deemed necessary by the STEP Core Team, alternative reporting procedures during fieldwork may be discussed.

## 2. Data Files Structure

**Standard 9.** The Data Files provided by the Survey Firm must include **all sampled households**, i.e.:

- all initially targeted households
- **and** all reserve households, **whether** they were activated or not.

The DEP should therefore start from the Sample File, which lists all initially targeted households and reserve households, and include all variables featured in the Sample File.

**Standard 10. All households listed in the Sample Files must feature in the data files with a final status code.**

The final status code will be recorded in a variable named **‘final status code’**, which will be equal to:

- one of the result codes listed in Module 11 of the Household Questionnaire for sampled households that yielded an individual interview (see Appendix 1),
- **OR** one of the reasons for replacing a household for sampled households that had to be replaced (see Appendix 2),
- **OR** ‘95’ for reserve households that were never activated.

In addition, an **‘activated household’** variable will be generated to record the ID of each household activated to replace another household. It should be equal to the ID of the activated reserve household if a household was replaced and to ‘99’ if the household was not replaced.

**Standard 11.** The Survey Firm will provide 2 data files to the STEP Consortium:

- (i) The first data file will provide data from the household roster (Module 1A), including information on all household members, in a **long** format (i.e., one row in the data file will correspond to one household member’s information);
- (ii) The second data file will provide data from Module 1A to Module 11 for all selected individual respondents, including information entered on the Cover Page and Back Cover.

**Standard 12.** The Survey Firm will provide data files in a STATA or SPSS format.

**Standard 13.** The DEP will conform to the STEP Variable Convention for naming, coding, and labeling variables. All names and labels will be entered in English.

The STEP Variable Convention will be provided to the Survey Firm.

**Standard 14.** For all questions including an “Other” option in response categories, a string variable will be generated to record respondents’ answers.

These answers will be translated into English in the final data files.

If an interviewer added an “Other” response that can be linked back to one of the initial codes, please recode the answer accordingly after consulting with the Project Manager.

If a particular answer features frequently in responses to a given question, the Survey Firm may decide to assign an additional code to the question for that answer. In such instances, an updated codebook will be provided to the STEP Consortium.

**Standard 15.** The ‘*Final Implementation Report*’ will include a section on “Data Quality”.

It will provide information on remaining errors or inconsistencies the Survey Firm was unable to rectify after revisiting or re-contacting respondents to check answers and correct errors. A summary based on Section 3 is provided in Appendix 3.

### 3. Consistency Checks

**The following checks must all be carried out during fieldwork to enable Interviewers and/or Supervisors to revisit respondents and correct inconsistencies.**

**The list of checks provided below is not meant to be exhaustive: it points to specific checks the Survey Firm should carry out, but any other error or inconsistency identified during fieldwork or data cleaning should be addressed as well.**

#### Check Digits

**Standard 16.** Survey Firms developing their own DEP must embed the STEP algorithm in their check digit validation program. This program is aimed at ensuring that the Household IDs entered in the data files are correct.

The STEP algorithm is presented in Appendix 4 of the *Operations’ Manual* and is included in the present document in Appendix 4.

🔗 *During fieldwork:* Any error should be immediately reported to the Project Manager and checked with Supervisors and Interviewers in the field.

## Activation of Reserve Households

**Standard 17.** The Survey Firm will check that, in each PSU, reserve households are activated following the sequential order set in the Sample Files and Sample Tracking Forms.

🔗 *During fieldwork:* Any error should be immediately reported to the Project Manager and checked with Supervisors and Interviewers in the field.

🔗 *Final Implementation Report:* Any error will be listed and fully documented in the ‘Final Implementation Report’ (e.g., reason for not complying with the activation order, grounds on which reserve households were selected).

## Checks Upon Closing a PSU

**Standard 18.** Upon closing a PSU, the Survey Firm will check the following:

- **All reserve households must have been activated according to the sequence** provided in the Sample Files;
- **All initial households from the Sample File must be accounted for**  
I.e., they must EITHER have led to a full interview and feature a result code in Question 7 of Module 11 of the Household Questionnaire,  
OR have been replaced by a reserve household, in which case they should feature a reason for replacement in variable ‘final\_status\_code’;
- **All activated reserve households must be accounted for**  
I.e., if a reserve household is listed as having replaced another one, it must EITHER have led to a full interview and feature a result code in Question 7 of Module 11 of the Household Questionnaire;  
OR have been replaced by another reserve household, in which case it should feature a reason for replacement in variable ‘final\_status\_code’;
- **All non-activated reserve households should feature a ‘final status code’ equal to ‘95’.**

🔗 *During fieldwork:* The Survey Firm should correct all errors during fieldwork, by checking against the original Sample Tracking Forms and exchanging with Supervisors.

If any error remains after checks in the field, the Survey Firm should immediately notify the STEP Consortium.

🔗 *Final Implementation Report:* Any remaining error should be listed in the ‘Final Implementation Report’.

## General Checks on the Household Questionnaire Data

**Standard 19.** The Survey Firm will check **all skips** included in the Household Questionnaire.

🔗 *During fieldwork:* Any skip error should be immediately reported to the Project Manager and checked with Supervisors and Interviewers in the field.

🔗 *Final Implementation Report:* The frequency of remaining skip errors should be listed in the ‘Final Implementation Report’ Module by Module.

**Standard 20.** The Survey Firm will check the **range** of each variable and correct all errors.

🔗 *During fieldwork:* Any range error should be immediately reported to the Project Manager and checked with Supervisors and Interviewers in the field.

## Cover Page

**Standard 21.** The Survey Firm will check that the **random numbers** from the Cover Page sticker that are entered in the data files are:

- (i) all integers ranging from 1 to 15
- (ii) **AND** all different from one another.

🔗 *During fieldwork:* In the event a sticker is found to be faulty, i.e., featuring erroneous random numbers, the Survey Firm will **immediately** inform the STEP Consortium.

🔗 *Final Implementation Report:* If any, the list of households whose random numbers are erroneous will be provided in the ‘*Final Implementation Report*’.

## Module 1A

**Standard 22.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct the consistency between household members’ age and date of birth (Questions 4 and 5).

🔗 *During fieldwork:* All identified inconsistencies will be checked and corrected during fieldwork.

🔗 *Final Implementation Report:* If any, remaining errors or inconsistencies pertaining to the above will be listed in the ‘*Final Implementation Report*’.

**Standard 23.** The Survey Firm will pay particular attention to checking the following:

- **Eligibility** of each household member, i.e., that each eligible household member was given a sequential number in Question 20 and conversely that no non-eligible household member was attributed one. An eligible household member is an individual aged 15 to 64 and deemed mentally capable;
- **Sequential numbers** of household members, i.e., that they are sequential integers (no duplicates, no skips).

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers.

🔗 *Final Implementation Report:* If any, the full list of households and household members featuring an error in Question 20 (Eligibility) will be provided in the ‘*Final Implementation Report*’.

## Module 1C

**Standard 24.** The Survey Firm will specifically check and correct inconsistencies pertaining to the following in Module 1C:

- **Number A** must be equal to the number of eligible household members;
- **Number B** must be equal to the first number from the random sticker on the Cover Page that is less or equal to Number A;
- The selected respondent's sequential number in Question 20 Module 1A must be equal to Number B in Module 1C.

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers.

🔗 *Final Implementation Report:* If any, the full list of households for which Number A and/or Number B were incorrectly determined and the full list of incorrectly selected household members by the Interviewer will be provided in the 'Final Implementation Report'.

## Module 2

**Special attention should be devoted to checking skips in Module 2.**

**Standard 25.** The Survey Firm will specifically check skips in Module 2.

🔗 *During fieldwork:* Any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers.

🔗 *Final Implementation Report:* The 'Final Implementation Report' will summarize the frequency of skip errors in Module 2.

**Standard 26.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct whenever possible the following:

- Consistency of the age respondent started first grade (Question 7 in the Master Household Questionnaire), in particular outliers will be identified and systematically checked with respondents (the highest and lowest 1% of the sample at the time of the data check);
- Consistency of the age respondent left formal education with the highest level of formal education completed (cross Questions 26 and 8B in the Master Household Questionnaire);
- If the respondent interrupted his/her studies, Interviewers must ensure respondents provide the number of years of interruption (Question 36 in the Master Household Questionnaire).

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any inconsistency or outlier will be communicated to the Field Manager. These will be systematically checked with respondents.

🔗 *Final Implementation Report:* Outliers and remaining inconsistencies, if any, will be documented in the 'Final Implementation Report'.

## Module 3

**Standard 27.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct the following in Module 3:

- Consistency with the measuring unit in Questions 2 and 3 (respondent's height and weight);
- Outliers in Questions 2 and 3 (the highest and lowest 1% of the sample at the time of the data check);
- Range of the number of days in Question 6 (should be less than or equal to 28 days).

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any inconsistency or outlier will be communicated to the Field Manager. These will be systematically checked with respondents.

🔗 *Final Implementation Report:* Outliers and remaining inconsistencies, if any, will be documented in the 'Final Implementation Report'.

**Standard 28.** Supervisors must check the self-reported height of respondents in 10% of the cases, by measuring respondents' height. The data thus collected will be entered by the data entry operators in the following variable '**height\_supervisor**'.

The Survey Firm will check the consistency of this variable with the self-reported height provided in Question 2.

🔗 *Final Implementation Report:* The full list of height inconsistencies will be provided in the 'Final Implementation Report'.

## Module 4A

**Special attention should be devoted to checking skips in Module 4A.**

**Standard 29.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct, whenever possible, the following:

- Skips in Questions 1 to 9;
- Range and correct coding in Question 19;
- Ranges and correct coding in Question 20 and 30;
- Range and correct coding in Question 25 and coding of "Other" response category.

🔗 *During fieldwork:* The Survey Firm will specifically check skips in Module 4A during fieldwork. Any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers, who should contact respondents again to correct the error.

🔗 *Final Implementation Report:* The frequency of remaining skip errors in Questions 1 to 9 will be provided in the 'Final Implementation Report'.

## Module 4B

**Standard 30.** Module 4B should only be applied to respondents who answered ‘yes’ to the following Questions in Module 4A:

- Question 1
- Or Question 2
- Or Question 3
- Or Question 4
- Or Question 5 and Question 6.

🔗 *During fieldwork:* The Survey Firm will specifically check this during fieldwork and any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers, who should contact respondents again to correct the mistake.

🔗 *Final Implementation Report:* The full list of respondents who were either erroneously excluded from Module 4B or erroneously interviewed in Module 4B will be provided in the ‘Final Implementation Report’.

**Standard 31.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct, whenever possible, the following:

- Range and correct coding in Questions 1, 4, and 11;
- Range in Question 3;
- Tasks listed in English in Questions 2 and 12;
- Ranges in Questions 5, 6, and 8 (days in the past week, hours per day, weeks per year);
- Consistency check in Question 7 (should be equal to Question 5 x Question 6).

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any inconsistency will be communicated to the Field Manager. Such instances will be systematically checked with respondents.

🔗 *Final Implementation Report:* Remaining inconsistencies, if any, will be documented in the ‘Final Implementation Report’.

## Module 4C

**Standard 32.** Module 4C should only be applied to respondents who answered Module 4B.

🔗 *During fieldwork:* The Survey Firm will specifically check this during fieldwork. Any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers, who should contact respondents again to correct the mistake.

🔗 *Final Implementation Report:* The full list of respondents who were either erroneously excluded from Module 4C or erroneously interviewed in Module 4C will be provided in the ‘Final Implementation Report’.

**Standard 33.** The Survey Firm will check that the occupation code provided in Question 1:

- Corresponds to one of the occupations listed by the respondent in Question 1 of Module 4B;
- Is the occupation with the highest number of hours in Question 7 of Module 4B or the highest number of weeks in Question 8 of Module 4B.

🔗 *During fieldwork:* The Survey Firm will specifically check this during fieldwork and any error will be communicated to the Field Manager to provide feedback to Supervisors and Interviewers, who should contact respondents again to correct the mistake.

🔗 *Final Implementation Report:* The full list of respondents whose occupation was not listed in Question 1 of Module 4B or is not the occupation with the highest number of hours or weeks (Questions 7 or 8 in Module 4B), after correction in the field, will be provided in the ‘Final Implementation Report’.

**Standard 34.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct the following in Module 4C:

- Time period is provided in Question 9;
- Outliers in Questions 24 and 26 (earnings), defined as the highest and lowest 1% of the observations at the time of data check;
- Pay period is provided in Questions 25 and 27, including “Other” is specified in English whenever relevant.

🔗 *During fieldwork:* The Survey Firm will carry out these checks during fieldwork and any missing time period or outlier will be communicated to the Field Manager. These will be systematically checked with respondents.

🔗 *Final Implementation Report:* The ‘Final Implementation Report’ will provide the following:

- Cross-tabulation between Question 6 of Module 4C and Question 3 of Module 4B (type of employment);
- Outliers in Questions 24 and 26 (earnings), defined as the highest and lowest 1% of the observations at the time of data check;
- List of missing values for pay period in Questions 25 and 27, including “Other” specification, when earnings are reported in Questions 24 or 26.

## Module 5A

**Standard 35.** Questions 1 to 20 in Module 5A should only be applied to respondents:

- Who answered Module 4C (i.e., worked in the past 7 days)
- OR worked in the past 12 months (the date must be checked based on the answer provided in Question 19 of Module 4A).

🔗 *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5A whenever required).

🔗 *Final Implementation Report:* The full list of respondents who were either erroneously excluded from or erroneously included in this section of Module 5A will be provided in the ‘Final Implementation Report’.

**Standard 36.** The Survey Firm will check that the occupation code provided in Question 1 is the same as the one provided in Question 1 of Module 4C.

🔗 *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5A whenever required).

🔗 *Final Implementation Report:* Remaining errors will all be listed in the ‘Final Implementation Report’.

**Standard 37.** The Survey Firm will check that the occupation code provided in Question 2 is the same as the one provided in Question 20 of Module 4A.

🔗 *During fieldwork:* Supervisors will have checked all erroneous cases with respondents (and reapplied Modules 5A whenever required).

🔗 *Final Implementation Report:* Remaining errors will all be listed in the ‘Final Implementation Report’.

**Standard 38.** Questions 21 to 28 in Module 5A should only be applied to respondents who haven’t worked in the past 12 months.

🔗 *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5A whenever required).

🔗 *Final Implementation Report:* The full list of respondents who were either erroneously excluded from or erroneously included in this section of Module 5A will be provided in the ‘Final Implementation Report’.

**Special attention should be devoted to checking skips in Module 5A.**

**Standard 39.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct the following in Module 5A:

- Skip errors from Questions 4 to 8 (‘Respondent does not read at work’) and from Questions 12 to 15 (‘Respondent does not write at work’);
- Skip errors from Question 2 to 21 (‘Respondents has not worked in the past 12 months’).

🔗 *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5A whenever required).

🔗 *Final Implementation Report:* The frequency of skip errors in these questions will be provided in the ‘Final Implementation Report’.

## Module 5B

**Standard 40.** Questions 1 to 26 in Module 5B should only be applied to respondents who worked in the past 12 months, i.e., who answered ‘yes’ either to Question 1 or to Question 2 in Module 5A.

🔗 *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5B whenever required).

🔗 *Final Implementation Report:* The full list of respondents who were either erroneously excluded from or erroneously included in this section of Module 5B will be provided in the ‘Final Implementation Report’.

**Standard 41.** Questions 27 to 36 in Module 5B should be applied to **all** respondents.

↳ *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Modules 5B whenever required).

↳ *Final Implementation Report:* The full list of respondents who were either erroneously excluded from or erroneously included in this section of Module 5B will be provided in the ‘Final Implementation Report’.

**Standard 42.** The Survey Firm will check that the occupation code provided in Question 1 is the same as the one provided in Question 1 or Question 2 of Module 5A.

↳ *During fieldwork:* Supervisors will check all erroneous cases with respondents (and reapply Questions 1 to 26 of Modules 5B whenever required).

↳ *Final Implementation Report:* Remaining errors will all be listed in the ‘Final Implementation Report’.

## Module 6B

**Special attention should be devoted to checking skips in Module 6B.**

**Standard 43.** The Survey Firm will specifically check and correct all skip errors in Module 6B.

↳ *During fieldwork:* Skip errors in Module 6B will be reported to the Field Manager and the Module will be reapplied to the respondent whenever possible.

↳ *Final Implementation Report:* Remaining errors will all be listed in the ‘Final Implementation Report’.

## Module 7

**Standard 44.** In addition to general checks mentioned at the beginning of Section 3, the Survey Firm will specifically check and correct the following in Module 7:

- Consistency between answers provided in Questions 7 to 11 and the Household Roster in Module 1A;
- Consistency between the ID provided by the respondent for his/her mother and father if they live in the same household (Questions 12 and 15) and the Household Roster (Module 1A), this will include checking that gender and age differences between the respondent and his/her parents are coherent and documenting all unusual circumstances.

↳ *During fieldwork:* All inconsistencies will be communicated to the field so Supervisors and Interviewers can check such instances with respondents.

↳ *Final Implementation Report:* After data cleaning, all remaining errors or inconsistencies pertaining to the above will be listed in the ‘Final Implementation Report’. In particular, all cases featuring inconsistent parent IDs will be listed and documented in the ‘Final Implementation Report’.

## Module 9

**Standard 45.** The General Booklet ID and the Exercise Booklet ID (for countries undertaking the Full Literacy Assessment) must feature in Module 9 and conform to the IDs provided in the Sample Tracking Forms.

🔗 *During fieldwork:* Any missing Booklet IDs must be immediately brought to the attention of the Field Manager.

🔗 *Final Implementation Report:* Any issue regarding missing Booklet IDs must be reported in the ‘*Final Implementation Report*’.

## Module 11

**Standard 46.** The IDs of the respondents who answered Module 1A and Module 1B must feature in Module 11, Question 3.

🔗 *During fieldwork:* Missing respondent IDs must be flagged during fieldwork to provide feedback to the field.

🔗 *Final Implementation Report:* If any, missing respondent IDs for Modules 1A or 1B must be listed in the ‘*Final Implementation Report*’.

**Standard 47.** Respondents must all have been attributed a Result Code in Question 7. The Result Code must comply with the predefined codes provided in the Household Questionnaire and be consistent with the questionnaire (Appendix 1).

## 4. Appendix

### Appendix 1. Final Result Codes – Household Interviews

#### 1. Full Literacy Assessment

Code	Result Code
<b>Household Module Begun, But Not Completed</b>	
21	Household <u>refused to continue</u> (time constraints, did not want the bother, other general refusal)
22	Household module could not be continued because of <u>unusual circumstance</u>
<b>Individual Modules 2-7 Not Begun</b>	
31	No household member in the <u>eligible range of 15-64</u>
32	Selected individual <u>refused</u> (time constraints, did not want to bother, other general refusal)
33	Selected individual could not be interviewed because of <u>unusual circumstance</u> (death in family, serious illness, fire in dwelling, etc.)
34	<u>Another household member refused</u> to let selected individual do individual modules
35	Selected individual <u>could not be contacted</u> after three visits during field period
36	Selected individual will be <u>absent for the entire field period</u> (information from household member)
37	Selected individual is <u>deaf</u> or so hard-of-hearing that cannot do survey, and no translator available
38	Selected individual has a <u>learning or mental disability</u> so that he or she is unable to answer (including emotional conditions like severe depression)
39	Selected individual has a <u>speech impairment</u> that prevents doing the survey, and no translator available
40	<u>Language problem</u> - selected individual did not speak a language that could be understood by survey team, and no translator available SPECIFY LANGUAGE_____
<b>Individual Modules 2-7 Begun, But Not Completed</b>	
51	<u>Individual refused to continue</u> (time constraints, did not want the bother, other general refusal)
52	<u>Individual interview could not be continued because of unusual circumstance</u>
<b>General Booklet (Module 9) Not Begun</b>	
61	Individual <u>does not read or write the language</u> of the General Booklet so refused to begin.
62	Individual <u>refused</u> to begin Booklet (time constraints, did not want to bother, other general refusal)
63	Individual could not do General Booklet because of <u>unusual circumstance</u>
64	Individual could not do General Booklet because of <u>blindness</u> or visual impairment.
65	Individual could not do General Booklet because of other <u>physical disability</u> (cannot hold pen)
<b>General Booklet (Module 9) Looked Through or Attempted, But Not Completed or Passed</b>	

71	Individual looked at Booklet but unable to read and write the language of the Booklet so did not begin
72	Individual began General Booklet but refused to continue
73	Individual had too long an interruption in the General Booklet, so it could not be resumed
74	General Booklet stopped because of unusual circumstance
75	Individual attempted the General Booklet but <u>did not mark</u> any answers
76	Individual did part or all of General Booklet but <u>did not pass</u> to Exercise Booklet (pass 3/8 in Section B)
<b>Exercise Booklet (Module 9)</b>	
81	Individual <u>did not attempt</u> any of the Exercise Booklet/ refused to start
82	Individual <u>had too long an interruption</u> in the Exercise Booklet, so it could not be resumed
83	Exercise Booklet stopped because of <u>unusual circumstance</u>
84	Individual <u>attempted some</u> of the Exercise Booklet but <u>did not mark</u> any answers
85	Individual <u>attempted the whole</u> Exercise Booklet but <u>did not mark</u> any answers
86	Individual <u>marked some</u> of the Exercise Booklet but <u>did not attempt all</u> the items
87	Individual <u>marked some items</u> of the Exercise Booklet and <u>attempted all</u> items
88	Individual <u>marked all</u> the questions of the Exercise Booklet

## 2. Partial Literacy Assessment

Code	Result Code
<b>Household Module Begun, But Not Completed</b>	
21	Household <u>refused to continue</u> (time constraints, did not want the bother, other general refusal)
22	Household module could not be continued because of <u>unusual circumstance</u>
<b>Individual Modules 2-7 Not Begun</b>	
31	No household member in the <u>eligible range of 15-64</u>
32	Selected individual <u>refused</u> (time constraints, did not want to bother, other general refusal)
33	Selected individual could not be interviewed because of <u>unusual circumstance</u> (death in family, serious illness, fire in dwelling, etc.)
34	<u>Another household member refused</u> to let selected individual do individual modules
35	Selected individual <u>could not be contacted</u> after three visits during field period
36	Selected individual will be <u>absent for the entire field period</u> (information from household member)
37	Selected individual is <u>deaf</u> or so hard-of-hearing that cannot do survey, and no translator available
38	Selected individual has a <u>learning or mental disability</u> so that he or she is unable to answer (including emotional conditions like severe depression)
39	Selected individual has a <u>speech impairment</u> that prevents doing the survey, and no translator available
40	<u>Language problem</u> - selected individual did not speak a language that could be understood by survey team, and no translator available SPECIFY LANGUAGE _____
<b>Individual Modules 2-7 Begun, But Not Completed</b>	

51	Individual <u>refused to continue</u> (time constraints, did not want the bother, other general refusal)
52	Individual interview could not be continued because of <u>unusual circumstance</u>
<b>General Booklet (Module 9) Not Begun</b>	
61	Individual <u>does not read or write the language</u> of the General Booklet so refused to begin
62	Individual <u>refused</u> to begin Booklet (time constraints, did not want to bother, other general refusal)
63	Individual could not do General Booklet because of <u>unusual circumstance</u>
64	Individual could not do General Booklet because of <u>blindness</u> or visual impairment
65	Individual could not do General Booklet because of other <u>physical disability</u> (cannot hold pen)
<b>General Booklet (Module 9) Looked Through or Attempted, But Not Completed or Passed</b>	
71	Individual looked through the Booklet but <u>unable to read and write the language of the Booklet so did not begin</u>
72	Individual began General Booklet but <u>refused to continue</u>
73	Individual had <u>too long an interruption</u> in the General Booklet, so it could not be resumed
74	General Booklet stopped because of <u>unusual circumstance</u>
75	Individual attempted the General Booklet but <u>did not mark any answers</u>
76	Individual <u>marked some items</u> of the General Booklet <u>but did not attempt all items</u>
77	Individual <u>marked some</u> of the General Booklet and <u>attempted all</u> the items
78	Individual <u>marked all questions</u> of the General Booklet

## Appendix 2. Final Result Codes – Replaced Households

01	Household refused to be interviewed (time constraints, did not want the bother, other general refusal)
02	Household refused or could not be interviewed because of unusual circumstances (death in family, illness, fire in dwelling, etc)
03	No knowledgeable household member could be found after 3 revisits (only child, non-competent adult, etc)
04	Temporarily absent / unavailable for field period (information from others)
05	No competent household member to interview (because of severe illness, mental disability, etc)
06	Language problem – no one in the household spoke a language that could be understood by the Survey Team and no translator was available Specify language _____
07	Dwelling could not be found / address is incorrect
08	Dwelling is empty
09	Dwelling is no longer habitable / dwelling is destroyed / dwelling has been converted to commercial use

### Appendix 3. Section on Data Quality in the Final Implementation Report

The ‘*Final Implementation Report*’ should include a section on “Data Quality”.

This section should include the following information, in addition to any other issue or comment the Survey Firm sees fit to add:

- Full list of cases presenting a check digit mismatch and action taken by the Survey Firm;
- Full list of cases in which the sequential order set by the Sample Files for activating reserve households was not respected, along with the reasons for not complying with the activation order and grounds on which reserve households were selected;
- Full list of unaccounted for initial and activated reserve households;
- Frequency of skip errors, module by module;
- Identified outliers that were confirmed with respondents;
- Full list of errors that occurred during the selection of individual respondents:
  - Full list of random number errors (‘N’ Numbers on the Cover Page sticker);
  - Full list of households in which an eligible household member was not given a sequential number in Question 20 Module 1A or conversely in which a non-eligible household member was attributed one;
  - Full list of households in which the sequential numbering of eligible household members was erroneous (e.g., duplicates, skips);
- The full list of households for which Number A and/or Number B were incorrectly determined in Module 1C;
- The full list of incorrectly selected household members (i.e., erroneous ID of the respondent based on Number B, Question 20 and household members’ IDs);
- The frequency of skip errors in Module 2;
- The full list of inconsistencies regarding the age at which respondents started first grade (Question 7 in the Master Household Questionnaire), in particular outliers will be documented (the highest and lowest 1% of the sample at the time of the data check);
- The full list of inconsistencies regarding the age at which respondents left formal education with the highest level of formal education completed (cross Questions 26 and 8B in the Master Household Questionnaire), in particular outliers will be documented (the highest and lowest 1% of the sample at the time of the data check);
- The frequency of missing values in the number of years of interruption for respondents who interrupted their studies (Question 36 in the Master Household Questionnaire);
- The full list of outliers in Questions 2 and 3 (the highest and lowest 1% of the sample at the time of the data check);
- The full list of inconsistent height measurements based on Supervisors’ revisits;
- The frequency of skip errors in Questions 1 to 9 in Module 4A;
- The full list of respondents who were either erroneously excluded from Module 4B or erroneously interviewed in Module 4B;
- The full list of respondents who were either erroneously excluded from Module 4C or erroneously interviewed in Module 4C;
- Cross-tabulation between Question 6 of Module 4C and Question 3 of Module 4B;
- Outliers in Questions 24 and 26 (earnings) in Module 4C;

- The full list of missing values for pay period in Questions 25 and 27, including “Other” specification, when earnings are reported in Questions 24 or 26 in Module 4C;
- The full list of respondents who were either erroneously excluded from or erroneously included in the section of Module 5A running from Question 1 to Question 20;
- The full list of cases in which the occupation code provided in Question 1 isn’t the same as the one provided in Question 1 of Module 4C;
- The full list of cases in which the occupation code provided in Question 2 isn’t the same as the one provided in Question 20 of Module 4A;
- The full list of respondents to which Questions 21 to 28 in Module 5A were applied although they hadn’t worked in the past 12 months;
- The frequency of skip errors from Questions 4 to 8 (‘Respondent does not read at work’) and from Questions 12 to 15 (‘Respondent does not write at work’) in Module 5A;
- The frequency of skip errors from Question 2 to 21 (‘Respondents has not worked in the past 12 months’) in Module 5A;
- The full list of respondents who were either erroneously excluded from or erroneously included in the section of Module 5B running from Question 1 to 26;
- The full list of respondents erroneously excluded from the section of Module 5B running from Question 27 to 36;
- The full list of cases in which the occupation code provided in Question 1 of Module 5B isn’t the same as the one provided in Question 1 or Question 2 of Module 5A;
- The frequency of skip errors in Module 6B;
- The full list of cases in which the ID provided by the respondent for his/her mother and father if they live in the same household (Module 7) and the Household Roster (Module 1A) are inconsistent;
- The frequency of missing values or ‘Don’t know’ answers will be provided for Questions 13 and 16 in Module 7;
- If any, the full list of issues regarding missing Booklet IDs in Module 9;
- If any, the full list of missing respondent IDs for Modules 1A or 1B;
- If any, the full list of issues in the attribution of result codes in Question 7 Module 11.

## **Appendix 4. Check Digits for Operational IDs**

### **Introduction**

The purpose of this document is to describe and standardize the computation and usage of check digits for the operational identification numbers used in STEP. It covers the following topics:

- Rationale for check digits and affected operational IDs
- Algorithm for the computation of STEP check digits and examples
- Microsoft Excel example and VBA syntax for the computation and validation of check digits and operational IDs

### **Rationale for check digits and affected operational IDs**

STEP uses certain, well-formatted operational IDs (i.e., household operational identification number). There is a strong need to reliably use these unique identification variables in the context of the various materials and data sources in order to create a correctly matched database. It is important to ensure that identification numbers are transmitted correctly, and in certain cases, not forged. This is especially true for any occasions where the ID number will be manually entered in a software system or recorded on paper materials. In the case of STEP, the person identification number will be manually entered or recorded in multiple occasions

Errors can occur, for example, when digits are written or typed in the wrong order, digits are written illegibly, one or more digits are mistyped, or one or more digits are omitted. The STEP Consortium therefore agreed to utilize check digits, i.e. an additional digit at the end of any operational ID that verifies the consistency of the entered or recorded ID. The implementation of the check digit will identify data capture errors immediately and facilitate the correction of either the ID entered or recorded on any paper materials.

The use of the check digit algorithm for person IDs and task instrument serial numbers as described in the next section is mandatory for all countries. Given various constraints and the need for standardized software, no other check digit algorithm will be approved or supported by the STEP Consortium.

The check digit computation is a country responsibility as part of the sample selection and creation of the survey control file (and specified in the Technical Standards). Consequently, the algorithm and software examples are provided in this note. The STEP Consortium will be responsible for implementing the necessary validation logic into the data management software and any subsequent data processing system. If the Firm is developing its own DEP, they have to embed in it the check digit validation logic.

### **Algorithm for the computation of STEP check digits and examples**

While many different check digit algorithms exist, the STEP Consortium settled with a derivation of one of the most commonly used check digit algorithms used today, the Universal Product Code (UPC). While this algorithm does not safeguard against each

and every possible error during capturing an ID (though it will most likely detect >99% of them), the algorithm is straightforward to understand and implement in common computer languages (Excel examples and VBA syntax will be provided below).

It should be noted that the original variants of UPC expect ID numbers of a fixed length (i.e., 11 digits in UPC-A). The STEP algorithm will work with any length.

The mandatory data type for all STEP operational IDs is numeric (i.e., no characters allowed) and no leading zeros must be used to pad IDs to the full length (i.e., a 10-digit ID should always have exactly 10 digits and start with numbers (1-9). With the additional check digit, the maximum allowed length of person ID is 12 digits.

In STEP case, the unique household ID will be a 5-digits number obtained by adding

$$\text{Household ID} = \text{PSU Sample ID} * 100 + \text{Household Sample ID}$$

Using this household ID, the algorithm to compute the STEP check digits is as follows:

- Step 1: Add the digits in the odd-numbered positions (first, third, fifth), starting from the left, together and multiply by three.
- Step 2: Add the digits in the even-numbered positions (second, fourth), starting from the left, to the result.
- Step 3: Find the result modulo 10 (i.e. the remainder when the result is divided by 10).
- Step 4a: If the result is zero, zero is the check digit
- Step 4b: If the result is not zero, subtract the result from ten and use the difference as the check digit.

The following example will use a fictional person ID (12345) to illustrate the check digit calculation.

- Step 1: The sum of all digits in odd-numbered positions is  $1+3+5=9$ , multiplied by three  $9*3=27$
- Step 2: The sum of all digits in even-numbered positions is  $2+4=6$ , added to the result in step 1 equals  $27+6=33$
- Step 3: The result modulo 10, i.e.  $33(\text{mod } 10)$ , is 3.
- Step 4b: The difference, i.e.  $10-3$ , is 7 and this is used as the check digit.

### **Microsoft Excel example and VBA syntax for the computation and validation of check digits and operational IDs**

This Annex is accompanied by the Microsoft Excel Workbook file “Check digit examples.xls” that includes two macros to calculate and verify check digits and a set of examples including the two used above. The macros are written in Visual Basic for Applications (VBA) and stored in the Excel module "PIAACFunctions" (press ALT-F11 to open the VBA editor or use the “Tools → Macro → Visual Basic Editor” menu).

- **CalculateCD:** Takes a string ID of flexible length  $\geq 1$  and computes the check digit

**Figure 1. VBA Function “CalculateCD”**

```
Public Function CalculateCD(ByVal ID As String) As Integer

Dim i As Integer 'integer for looping through the ID string
Dim intIDLength As Integer 'integer to hold the length of the ID
string
Dim intSumOdd As Integer 'sum of all digits in odd-numbered
positions
Dim intSumEven As Integer 'sum of all digits in even-numbered
positions
Dim intSum As Integer 'total sum
Dim intSumMod10 As Integer 'sum modulo 10
Dim intCheckDigit As Integer 'integer check digit

    'check if the ID passed is numeric, otherwise quit with
warning
    If Not IsNumeric(ID) Then
        MsgBox "ID is not numeric", vbCritical + vbOKOnly
        Exit Function
    End If

    'get ID length
    intIDLength = Len(ID)

    'loop through the ID, starting from the left
    For i = 1 To intIDLength
        'conditional computation for odd and even-numbered
positions
        If (i Mod 2) = 1 Then
            'add to the sum of odd-numbered digits
            intSumOdd = intSumOdd + CInt(Mid(ID, i, 1))
        Else
            'add to the sum of even-numbered digits
            intSumEven = intSumEven + CInt(Mid(ID, i, 1))
        End If
    Next

    'multiple odd-numbered sum by three and add even-numbered sum
    intSum = intSumOdd * 3 + intSumEven

    'compute the modulo 10 remainder
    intSumMod10 = intSum Mod 10

    'conditional logic based on the remainder
    If intSumMod10 = 0 Then
        'if zero, use zero as the check digit
        intCheckDigit = 0
    Else
        'if not zero, use 10 minus the remainder as the check
digit
        intCheckDigit = 10 - intSumMod10
    End If

    CalculateCD = intCheckDigit

End Function
```

- **VerifyID:** Takes a string ID with check digit and returns TRUE if the ID is valid, otherwise FALSE. Note that VerifyID is dependent on Calculate

**Figure 2. VBA Function “VerifyID”**

```
Public Function VerifyID(ByVal IDwithCD As String) As Boolean

Dim strID As String 'string to hold the operational ID (i.e. w/o
check digit)
Dim intIDLength As Integer 'length of the ID w check digit
Dim intCheckDigit As Integer 'check digit passed to the function
Dim intCheckDigitVerify As Integer 'recalculated check digit

    'check if the ID passed is numeric, otherwise quit with
warning
    If Not IsNumeric(IDwithCD) Then
        MsgBox "ID is not numeric", vbCritical + vbOKOnly
        Exit Function
    End If

    'get length of ID w check digit
    intIDLength = Len(IDwithCD)

    'get operational part from ID w check digit
    strID = Left(IDwithCD, intIDLength - 1)

    'get check digit part from ID w check digit
    intCheckDigit = CInt(Right(IDwithCD, 1))

    'recalculate check digit
    intCheckDigitVerify = CalculateCD(strID)

    'compare passed and recalculated check digit and return TRUE
if both agree
    If intCheckDigit = intCheckDigitVerify Then VerifyID = True

End Function
```

As illustrated in columns C and F of the worksheet “Check Digit Examples”, the macros can be used in very much the same way as the built-in Excel functions.

More specifically, use the formula “=CalculateCD(A1)” to calculate a check digit for an operational ID using the cell reference to the operational ID as the argument (“A1” in the example). Provided that the argument is numeric, the function will return the check digit which can be appended to the operational ID (see column D for an example).

Use the formula “=VerifyID(A1)” to verify an operational ID including a check digit using the cell reference as the argument (“A1” in the example). The function will return TRUE for a valid ID and FALSE for an invalid one.

To work with the Excel example, you need to set your macro security settings in Excel to "medium" or below (menu “Tools → Macro → Security”) and also tell Excel that you want to "enable macros" when you open the file.

## **References**

The following resources may be helpful to further understand and implement check digit algorithms.

- Description of UPC on Wikipedia:  
[http://en.wikipedia.org/wiki/Universal\\_Product\\_Code](http://en.wikipedia.org/wiki/Universal_Product_Code)
- A nice introduction to the rationale, ease of use, strength, and flexibility of various algorithms: <http://www.academic.marist.edu/mwa/idsn.htm>