Feed the Future

Survey Implementation

Document

Data Processing and

Finalization Procedures for the

Feed the Future  
Zone of Influence Survey

July 2019

This document was produced for review by the United States Agency for International Development. It was prepared by ME&A and its sub-contractor ICF for Feed the Future Global Program Evaluation for Effectiveness and Learning Task Order (PEEL-TO), Contract Number GS-23F-8144H/AID-OAA-M-12-00006.

**Recommended Citation**

Genevieve Dupuis, Hanna Useem, and Kiersten Johnson, 2019.

*Feed the Future Zone of Influence Surveys: Data Processing and Finalization Procedures for the Feed the Future Zone of Influence Survey.*

Washington, DC: Bureau for Food Security, USAID

**Contact Information**

Feed the Future

1300 Pennsylvania Ave, NW

Washington, DC 20004

[www.feedthefuture,gov](about:blank)

**Table of Contents**

[Abbreviations i](#_heading=h.1fob9te)

[Introduction 1](#_heading=h.3znysh7)

[Purpose of the manual 1](#_heading=h.tyjcwt)

[Suggested file structure 1](#_heading=h.3dy6vkm)

[Objectives for finalizing the data 2](#_heading=h.1t3h5sf)

[A.](#_heading=h.4d34og8) Confirm count of data files 3

[B.](#_heading=h.2s8eyo1) Concatenate all cluster data files 4

[C.](#_heading=h.17dp8vu) Check count of questionnaires 4

[D.](#_heading=h.3rdcrjn) Review frequencies 5

[E.](#_heading=h.26in1rg) Review outliers 6

[F.](#_heading=h.lnxbz9) Complete secondary editing 7

[G.](#_heading=h.35nkun2) Review LandPKS database 7

[H.](#_heading=h.1ksv4uv) Process “other” answers 8

[I.](#_heading=h.44sinio) Calculate weights 9

[J.](#_heading=h.2jxsxqh) Recode multiple-response variables 10

[K.](#_heading=h.z337ya) Add analytical anthropometry variables 10

[L.](#_heading=h.3j2qqm3) Remove PII 10

[M.](#_heading=h.1y810tw) Export data 11

[N.](#_heading=h.4i7ojhp) Produce data release documentation 11

# Abbreviations

ICDM In-Country Data Manager

PII personably identifiable information

USAID United States Agency for International Development

ZOI Zone of Influence

# Introduction

Feed the Future seeks to reduce poverty, hunger, and undernutrition among women and children, and to increase income, women’s empowerment, dietary diversity, and appropriate feeding practices. Program efforts are concentrated in Zones of Influence (ZOIs) in Feed the Future target countries. Progress in achieving Feed the Future’s objectives is tracked using population-based performance indicators. The purpose of the Feed the Future ZOI Survey (2018–2019) is to provide U.S. Government interagency partners, the United States Agency for International Development (USAID) Bureau for Food Security, USAID Missions, host country governments, and development partners with information on the current status of the Feed the Future ZOI-level indicators. The survey is designed to monitor progress and determine whether there has been statistically significant change over time at the population level in key outcome and impact indicators, with the expectation that the effects of a project should spread beyond beneficiaries to the general population in the Feed the Future ZOI over the life of the project.

# Purpose of the manual

This manual outlines the procedures and guidelines for the review and finalization of the ZOI Surveys in Feed the Future countries and provides guidance for producing the datasets used in analysis. The [CONTRACTOR] Data Manager, CSPro Programmer, and the [SUBCONTRACTOR] In-Country Data Manager (ICDM) will work together to follow the procedures outlined in this document after the completion of field data collection and Central Office data processing done by the ICDM. The manual is an overview of the review, cleaning, and editing procedures required to produce a clean dataset, and each step should be adapted to the country’s survey. This manual assumes that the cleaning and final processing of the data is carried out in-country, using the same computer or server on which Central Office data management occurred. However, it is possible to follow these procedures on a separate computer (e.g., if the [CONTRACTOR] Data Manager is finalizing data from a remote location). The [CONTRACTOR] Data Manager will lead the procedures outlined in this manual.

# Suggested file structure

All data and programs used to process Feed the Future ZOI Survey data are stored on the Central Office computer (or server) and managed by the ICDM throughout fieldwork. Files will be stored in the root directory of the computer in a folder, “C:\*c*FTF,” where *c* is the first letter of the country for the survey. Subfolders in the C:\*c*FTF folder will hold the data for processing. In addition to the file structure required for data processing during fieldwork (see the [ICDM manual](https://drive.google.com/drive/u/1/folders/12krVrbzRGy6HSS2LlpIgxQWGFHrsr_S2)), below is the suggested file structure for final processing of the data files. Note that folders are suggested locations for programs to complete finalization procedures, but other folder structures may be used to complete processing of final datasets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Root Directory** | | |
|  | | C:\cFTF | | |
|  |  | \CENTRAL |  | All programs the ICDM uses to process, review, and finalize data |
|  |  | \LOOKUP |  | Files used to control data processing and track cluster status |
|  |  | \CODATA |  | All survey data for central processing, in subfolders: |
|  |  |  | .\AREA | Original plot area polygons received from the field and supervisor barcode databases |
|  |  |  | .\RECEIVED  .\ACCEPTED | Original survey data received from the field  Data that have passed basic structure checks and are ready for editing |
|  |  |  | .\FINAL | Final data after editing and finalization |
|  |  | \FINDATA  \LANDPKS  \PRELIM  \EXPORT |  | Complete survey datasets (concatenated for all clusters), at each stage of finalization  Programs for cleaning LandPKS database  Programs for processing and finalizing datasets  Programs for exporting final data into desired formats |

# STEPS for finalizing the data

The objectives and steps for the final data processing are to:

1. Create finalized datasets for the survey. The activities needed to do are as follows:
   1. Carry out initial checks to ensure that all data files from the field were received in the server/Central Office and completed initial processing and editing.
   2. Concatenate cluster data files into a full dataset for the survey.
   3. Carry out initial checks to ensure that the correct totals of clusters, households, and questionnaire modules have been collected.
   4. Check the frequency tabulations.
   5. Review outliers and unusual values.
   6. Complete outstanding secondary editing on the concatenated data file, and make adjustments to the cluster data files if needed.
   7. Check LandPKS database barcodes against the main data barcodes in the survey.
   8. Process and merge the “other” answers with the main data file.
   9. Calculate and add the sample weights to the main data file.
   10. Recode multiple-response questions.
   11. Add the World Health Organization Z-scores to the main data file.
   12. Remove personally identifiable information (PII) from main data file.
   13. Export data for analysis.
   14. Produce data release documentation.
2. Create SPSS/SAS/Stata/R data files for dissemination to outside organizations.
3. Create SPSS/SAS/Stata/R data files for internal use at the USAID Bureau for Food Security.
4. Document discrepancies, issues, or caveats with the final datasets that are relevant to any potential data users.

## Confirm count of data files

After all data from the field have been received in the Central Office and before finalization of the datasets can begin, all data files received must be accounted for. That is, for every cluster, all data collected in the field should be present in the appropriate locations. Throughout the course of fieldwork, as data collection is completed in each cluster, the data are reviewed and processed in the Central Office, led by the ICDM (see the [ICDM manual](https://drive.google.com/drive/u/1/folders/12krVrbzRGy6HSS2LlpIgxQWGFHrsr_S2)). All cluster data files must have passed all levels of Central Office processing (as described in the ICDM manual), and each version of the cluster survey data must be present in the appropriate locations. The required data files, their locations, and file naming conventions are described as follows, in accordance with the core Feed the Future ZOI guidance documents and programs.

The following data files must be present in the Central Office database (via transfer from the FTP server) in C:\*c*FTF\CODATA\RECEIVED (raw data from the field), C:\*c*FTF\CODATA\ACCEPTED (data that have passed initial structure checks), and C:\*c*FTF\CODATA\FINAL (data that have completed secondary editing):

|  |  |
| --- | --- |
| A*ccc*.dat | Questionnaire data files, containing all households and all modules for each cluster |
| G*ccc*.dat | GPS coordinate data files, containing GPS point for each household |
| O*ccc*.dat | Other answer data files, containing all “other” answers given in the cluster’s interviews |

where *ccc* represents the cluster number.

For every team, the following files should be present in the Central Office database (via transfer from the FTP server) in C:\*c*FTF\CODATA\AREA:

|  |  |
| --- | --- |
| T*nnn*.dat | Transmittal sheet database, containing all barcodes used to identify plots in every cluster, for each team |

where *nnn* represents the team number.

For each plot that was measured in the field, the following files should be present in the Central Office database (via transfer from the FTP server) in C:\*c*FTF\CODATA\AREA:

|  |  |
| --- | --- |
| GFAM-*x*.kml | Plot area polygons, containing one or more plot areas per file, identified by the assigned barcode |

where *x* represents the system-generated file name for polygon file.

Note that depending on frequency of transmission, area polygons are often grouped (e.g., all polygons mapped by an interviewer in a cluster may be present in one GFAM-*x*.kml file). Each polygon in the .kml file is identified by the barcode assigned to the specific plot, allowing for matching of polygons back to the datasets.

If any files are missing, data must be recovered and resubmitted from the tablet of the appropriate supervisor (and, if necessary, interviewers). Any survey data that were not fully processed in the Central Office during fieldwork must be completed. All files must be present and complete before the processing of the final, full datasets can begin. After all files are accounted for and prior to cleaning and finalization, a backup should be created of the ICDM’s directory (C:\*c*FTF\) to ensure that an original version of the data is preserved in a separate location from the processing.

## Concatenate all cluster data files

After all cluster data files are accounted for and all Central Office processing has been completed for each cluster (led by the ICDM during the course of fieldwork), concatenate the individual cluster data files (C:\*c*FTF\CODATA\FINAL\A*ccc*.dat) into one dataset, creating version one (**V1**) of the dataset. After collapsing the data files, sort the **V1** dataset by cluster and then by household number, creating version two (**V2**).

## Check count of questionnaires

Check that the number of questionnaires per cluster matches the number of questionnaires expected in the sample per cluster. If any clusters had fewer numbers of households—either due to sample design or potential field issues—this should be reflected in the total number of questionnaires in those clusters.

To check the count of questionnaires, the structure check application provided for Central Office processing (StructureCheck.ent in C:\*c*FTF\CENTRAL) that was used to initially check the structure of individual clusters may be used to check the structure of the whole dataset (**V2**). Alternatively, a close review of frequencies of **V2** can indicate the number of questionnaires per cluster for comparison to the expected sample.

If there are differences between the sample and the **V2** dataset, these differences must be inspected. Sometimes questionnaires are accidentally deleted during secondary editing, or files can become corrupted. If this is the case, missing questionnaires are likely to be recovered from several places, including the following folders (in order of highest level of data processing):

* C:\*c*FTF\CODATA\FINAL
* C:\*c*FTF\CODATA\ACCEPTED
* C:\*c*FTF\CODATA\RECEIVED
* Server locations
* Supervisor tablets (including backups)
* Interviewer tablets (including backups)

Any missing questionnaires that can be recovered must be added back into the appropriate cluster data file in C:\*c*FTF\CODATA\FINAL, and Step B must be repeated.

If there are discrepancies in the expected number of households per cluster as indicated by the Sampling Specialist and the actual number of households in each cluster, then these need to be investigated. Any differences between the numbers of households surveyed and the numbers of households in the sample design may require consultation with the Sampling Specialist and adjustment to the sample weighting. All differences should be documented in the Data Release documentation released with the final datasets.

## Review frequencies

After a count of questionnaires is validated, it is important to produce frequencies to check the **V2** dataset. Note that it is important to run the frequencies program at the beginning and end of final data processing, as well as in between stages of processing to catch any errors introduced during processing.

The following need to be checked while reviewing the frequencies:

**Unusual and outlier values**

Generally, unusual and outlier values are found for variables that represent numeric values, such as age, height, weight, land area, times, and numbers of items. Unusual values often indicate that data entry errors have been made during the interviews. Look for values that are outside the range of the majority of the responses, values with low or single frequencies, values that seem unusually precise, and values that seem unlikely in the context of the question (e.g., values of   
1–5 for the hour of the start/end of interview of an anthropometry measurement).

Any values of concern should be reviewed for any potential context clues that might indicate data entry error. To do so, see **E. Review outliers.** It is important that no changes are made to the data unless it is absolutely certain that the change is correct. If it is clear that a value is implausible, but a corrected value cannot be determined with certainty, an out-of-range code may be assigned (e.g., 7, 97, 997 for numeric values can represent “invalid response”). However, such values should be used very rarely because replacing data with an invalid code decreases the amount of information collected in the field.

In the instance of a computer-assisted personal interviewing survey, most reviews will not result in a change to the data because it will not be clear whether a data entry error was made. If data were collected on paper, all unusual and outlier values should be double-checked against the paper survey to ensure that the responses on paper align with the datasets.

**Missing or incorrect value labels**

Missing labels indicate that a response fell outside the range of possible values defined in the data dictionary for a variable. If frequencies are run in CSPro, a global search for the “@” character will reveal any response categories (value sets) that are missing labels. If frequencies are produced in another software, review the results to identify any missing labels.

**Default values**

Default values occur when a variable is not assigned a value in the data collection programs. This differs from a value of “not applicable” in CSPro; if a variable is not applicable (due to a skip in the questionnaire, for example), the variable is assigned a value of “not applicable,” which differs from lacking a value and being set as default. Default values should not appear in any of the variables. If they do occur, they can typically be identified with the following:

* There are stars in the CSPro data file (\*).
* There are blanks in the CSPro data file where the keyword <NotAppl> has not been specified as a special value in the dictionary.
* The value label “Default” appears in frequencies of the datasets.

In all cases, the data dictionary should be altered to ensure that special values are correctly handled, and the frequencies for the relevant variables should be re-run and checked to ensure that no remaining default values occur. It is also possible that a default value has been created during other stages of processing, during which a coding error occurred or a secondary edit was not properly handled (e.g., incorrect skip, or a modification in secondary editing in which the questionnaire was not properly exited and the modification affected the path through the questionnaire). The cause of any default value must be identified and resolved so that no default values remain.

**Skip patterns**

While reviewing the frequencies, check to ensure that the skip patterns of the questionnaire were followed correctly. If a skip pattern was not programmed correctly, write a batch program to either assign “missing” (code 9,99,999… for numeriv variables or “?” for string variables) or “not applicable” (notappl in CSPro) to the variables so the skip pattern is corrected. Any skip errors that are discovered must be documented in the Data Release documentation released with the final datasets.

**Spelling and grammar check**

In addition to reviewing missing labels, review frequencies for misspellings, grammar mistakes, or typos in the data dictionary—both for response categories and the variable labels. All issues should be corrected and cleaned to ensure that a clean dataset will be exported. In all cases, simply correcting the data dictionary in CSPro is adequate—new exports of the data will carry over the corrections.

## Review outliers

After reviewing the frequencies of **V2** and identifying variables with values for review (outliers, unusual values, etc.), write a batch program that provides a list and count of the outlier cases in the **V2** dataset. The output will first serve as a list of cases, as described above, for review. This program should output each case in which a certain variable meets the outlier criteria and include the cluster and household for reference.

For example, during data collection, Question 213 in Module 2 of the questionnaire (“How long does it take to go [to the water source], get water, and come back?”) allows values ranging from 1 minute up to 900 minutes. However, unusual values are commonly reported, such as times for fetching water of 72 minutes, 800 minutes, or 453 minutes. Although those examples are possible, they are unlikely, and it is important to review the data entry (and in this case, reviewing the times from other households in the cluster may provide context). A batch program may output the following information:

Cluster: 001 Household: 010 V213 = 72 minutes

Cluster: 012 Household: 004 V213 = 800 minutes

Cluster: 103 Household: 005 V213 = 453 minutes

With this information, review the questionnaire for Question 213 (variable name V213) in the corresponding cluster and household for each outlier or unusual response. Review as described above. No edit should be made to the data unless a correction is absolutely clear. Any required changes should be made to the individual cluster files (following the methods described in the [ICDM manual](https://drive.google.com/drive/u/1/folders/12krVrbzRGy6HSS2LlpIgxQWGFHrsr_S2)).

After reviewing and making necessary modifications to the data, re-run the outlier output, including the counts of each variable’s outliers. The variables with outlier or unusual values should be documented in the Data Release documentation released with the final datasets, so that data users are aware of special cases they may need to consider in their analysis.

## Complete secondary editing

During the course of fieldwork, the ICDM leads the secondary editing of data (see the [ICDM manual](https://drive.google.com/drive/u/1/folders/12krVrbzRGy6HSS2LlpIgxQWGFHrsr_S2)). Editors are instructed to only make edits to the data that they are certain of (following the secondary editing guidelines outlined in the ICDM manual), and any edits they are not sure of should be left for review during final processing. The secondary editing program used for cluster file editing should be used to do a more in-depth check on the dataset as a whole (**V2**). This program is located in the Central folder (C:\cCFTF\CENTRAL\SecondaryEditing.bch). Run the program on **V2** and review all remaining messages. If there are messages that need attention, go back to the cluster file and make the edit (through the Central Office menu). Review the total number of each message and prioritize according to which messages can be edited (i.e., where there is clear guidance in the secondary editing manual on how to modify the data) and which messages appear most frequently. After a full review of the secondary editing messages, repeat the processing steps, beginning with concatenation of the cluster data files.

After the review and resolution of secondary edits where appropriate, an accounting of remaining messages (i.e., those that cannot be resolved in the datasets) should be documented in the Data Release documentation released with the final datasets. A suggested format for reporting the remaining messages in the documentation is as follows:

* There are X women whose ages differed in the household roster and the female-WEAI module.
* There are Y children who were listed as 0–5 years in the household roster, but in the nutrition/anthropometry module had birth dates reported as older than 5 years old (and so were ultimately not eligible to be measured for anthropometry).

## Review LandPKS database

For the identification of plots in the LandPKS application, each is assigned a five-digit, anonymous barcode to protect the identities of farmers whose soil quality is measured using the LandPKS application. Barcodes are required because the data are loaded in the LandPKS website, through which any user can download the data collected by the app. To prevent matching of data from the website to individual respondents, an anonymous barcode must be used to name the plots. This barcode exists in the survey data and can be used to match LandPKS data that are downloaded to the survey data.

To download the LandPKS database, navigate to the LandPKS Data Portal (<https://landpotential.org/data-portal/>). In the “Recorder Email” field, enter the email address that was used during data collection. In the “Data Category” field, select the “LandInfo” option. Click on the blue “Export” button above the “Recorder Email” field. The collected data will be downloaded into an Excel file. If you need to view the metadata, select the “Metadata for LandInfo” option in the “Data Category” field, and export in the same way. This process will need to be repeated for each email address that was used to collect data in the field. Combine all downloads into an Excel file holding all the LandPKS data from all teams. The name of each plot is indicated with a barcode entered by the agricultural interviewer.

Ensure that all the barcodes in the survey dataset align with the barcodes identifying plots in the data downloaded from the LandPKS website. To do so, the list of plot barcodes downloaded from the LandPKS website should be compared to the list of plot barcodes in the **V2** dataset. If there are mismatches, or missing barcodes from either the LandPKS data or the survey data, the plot barcodes must be closely reviewed. Possible reasons for mismatches are as follows:

* Input error of plot barcode into the LandPKS application was made when naming a soil assessment.
  + Barcodes entered into the CSPro data collection programs are checked, making it essentially impossible to enter an incorrect barcode. However, the LandPKS application does not have the ability to check the validity of the plot barcode. Therefore, an error in the entry of the plot barcode into the LandPKS app would result in a mismatch.
* A farmer refused to allow the agriculture interviewer to assess the quality of the soil using the LandPKS application.
  + This would result in a plot barcode in the survey data that is missing from the LandPKS database. In these cases, the result of the LandPKS measurement in the survey database should be recorded as “refused” or “not measured.”

In many cases, barcodes that are incorrectly entered in the LandPKS application can be corrected based on a process of elimination. The correction can be triangulated using the supervisor’s transmittal database—the list of plot barcodes are entered into the supervisor’s tablet for each cluster before closing the cluster. If a group of barcodes that do match between LandPKS and the survey data are found, it can increase the likelihood of correcting a data entry error in the LandPKS database.

Any mismatches in the barcodes should be recorded in the survey database (result of the measurement should be changed to “not measured”).

## Process “other” answers

When respondents give an answer to a question that is not represented in the predetermined response options, and an “other” response category is available, the interviewer records the response given. The “other” answers are recorded and stored in the C:\*c*FTF\CODATA\*Occc*.dat data files. The “other” answers must be reviewed, and, if possible, recoded back into the predetermined response categories. Often, “other” answers consist of the predetermined response categories or synonyms for the response categories. The following steps describe a method for reviewing and recoding the “other” answers:

1. Concatenate the O*ccc*.dat files into one file, containing all “other” answers reported in the survey. Each “other” answer is identified by the cluster, household, variable, and occurrence (if applicable—for example, questions are asked of more than one child in the nutrition module, so the occurrence, or child, must be included in the identification of the “other” answer). The O*ccc*.dat files also contain a column with a 0 field at the end that can be used to recode the answers if possible.
2. Sort the file first by question number and then alphabetically by answer. This will group all “other” responses that are identical to help minimize the work of determining potential recoding.
3. Review all the “other” answers and determine whether they can be recoded into a current response category. If it is possible to recode, change the 0 field in the data file to the code representing the response category to which the “other” should be recoded. This can be accomplished easiest in an Excel format that would eventually be saved back into the required format for merging into the survey data, but this is not required (i.e., recoding can be done directly in the dataset). If the “other” answer cannot be recoded, leave the recode field as 0.
4. If there is an “other” answer that consistently appears for a variable, a new response category can be created. Add the category to the value set in the CSPro data dictionary and recode all “others” to the newly assigned value. A general rule is if the answer accounts for 10 percent or more of the total responses to a question, a new response category should be added to the value set.
5. After reviewing all the “other” answers, the dataset containing the recoded values should be merged into the main dataset, creating **V3** of the final dataset. Note that skip patterns may be affected after “other” answers have been recoded. When merging the “other” answers, it may be necessary to assign “notappl” (not applicable) or “missing” to some variables based on the change of the skip patterns due to the recoding of an “other” answer. To determine required changes from skip patterns, carefully review the questionnaire and skip patterns for questions in which “other” answers are recoded.
6. Review a new set of frequencies of the **V3** dataset for potential issues in recoding the “other” answers, following the guidance described in **D. Review frequencies**.

## Calculate weights

To allow the Sampling Specialist to calculate the weights required for the survey, export the numerators and denominators of the variables that will used for the weights, differentiated by cluster (because weights are assigned per cluster). The inputs for weights differ in each survey and are described in detail in the Sampling Manual. It is critical to do this step after all the “other” answers have been processed. The standard weights are for households, children 0–4 years, children 0–1 years, women 15–49 years (nutrition/anthropometry), primary decisionmakers (male and female), and farmers (overall and for each value chain commodity). Provide the numerators and denominators by cluster for the weights to the Sampling Specialist. The Sampling Specialist will return the weights.

Add the weights to the dataset, under the designated weight variables available in Module 0: Household Identification (record HH\_ID) of the data dictionary. The weights should be written into the available weight variables, according to the survey’s requirements. The standard weight variables are WGT\_HH, WGT\_MPDM, WGT\_FPDM, WGT\_C5, WGT\_C2, WGT\_W, WGT\_F, WGT\_V1, WGT\_V2, WGT\_V3, and WGT\_LV; add any other weights required for the survey. The addition of the weights will create **V4** of the dataset.

After the weights have been added to the data file, the weighted and unweighted frequency distributions should be produced. The total number of complete interviews for households, as well as each subgroup of respondents for which weights were calculated, should be the same for both weighted and unweighted distributions if the weights are correctly calculated. If they do not match, work with the Sampling Specialist and review all programs associated with the weight calculation to determine where the error lies.

## Recode multiple-response variables

For all multiple-response questions in the survey in which the respondent was able to give more than one answer (indicated in the questionnaire by “SELECT ALL THAT APPLY” on questions with response categories whose codes are letters), the string variables that hold the respondent’s original answer must also be recoded into numeric fields with value sets of 1-YES, 2-NO. For each possible response category, a new variable is created that indicates whether the response category was selected: if so, the new variable is code 1-YES; if not, the new variable is code 2-NO. For example, Question 217 in Module 2 (“What do you usually do to make the water safer to drink? Anything else?”) indicates that the interviewer should SELECT ALL THAT APPLY. Therefore, the dataset should include the original string variable response, as well as a new variable for each possible response. For examples and more detail, see the Core Codebook.

To recode multiple-response variables, each new variable corresponds to a response category from the original variable. Populate each new variable by searching the original variable for the corresponding code; if the category’s code is found, the new variable is coded 1-YES. Otherwise, code the new variable as 2-NO. Repeat this process for each response category to create each of the new variables. This must be done for each multiple-response question in the survey.

The process of recoding multiple-response questions will result in a new data dictionary, as well as **V5** of the final dataset.

## Add analytical anthropometry variables

For children, anthropometry data must be used to calculate the analytical Z-score variables for Feed the Future indicators, including wasting and stunting, among others. The variables that must be calculated and added to the dataset are as follows:

* HAZ: Height-to-age Z-score
* WAZ: Weight-to-age Z-score
* WHZ: Weight-to-height Z-score

Information on calculating these variables follows World Health Organization recommendations and is found at <https://www.who.int/nutgrowthdb/software/en/>.

Add the variables to the dataset, under the designated anthropometry variables available in Module 5A: Children’s Anthropometry (record ANTHRO\_CHILD) of the data dictionary. The variables should be written into the available anthropometry variables, according to the survey’s requirements. The standard variables are HAZ, WAZ, and WHZ. The addition of the anthropometry analytical variables will create **V6** of the dataset.

## Remove PII

All PII must be removed from the dataset to protect the anonymity of survey respondents. PII is any information that could be used to identify any respondents or households that were selected for the survey. This includes, but is not limited to, cluster and village names, addresses, and names (e.g., household and anthropometry rosters). Be sure to consult the Sampling Specialist to determine the lowest level of geographical identifiers that can be included in the dataset. In most cases, geographical identifiers are not allowed below Admin 2 for the country.

A suggested way to do this is to use CSPro’s Reformat Data feature. First, create a new version of the dictionary that has all the PII variables removed—delete them from the new dictionary. For example, V101 (Name of household member) must be deleted. Carefully review the entire dictionary for any PII and remove it from the new dictionary. Then, using CSPro’s Reformat Data feature, input the old dictionary and **V6** of the data, as well as the new dictionary. This will produce **V7** of the dataset that no longer includes PII.

## Export data

**V7** of the dataset must be exported for analysis and for data users. If the above steps are performed in another software, the export procedure should be done earlier in the finalization procedures.

To export CSPro data into the desired formats (including SAS, Stata, SPSS, R), use the export program provided with the core programs, located at: C:\*c*FTF\EXPORT\Export.bch. This program creates links between the modules required for export formats other than CSPro. Ensure that the appropriate dictionary is used for the version of the dataset that is being exported.

Note: It is also possible to export data using CSPro’s Export Data feature, which may allow the export to be customized or be done for smaller, incremental exports. This may aid in the various steps of reviewing, processing, and finalizing, but ultimately, the final dataset is export should be done using the provided export program because it is aligned with the Guide to Statistics for analysis.

## Produce data release documentation

Finally, as referenced throughout this manual, with the release of the final datasets, the datasets should be accompanied by Data Release documentation. This document should include any and all caveats, notes, issues, and outstanding inconsistencies in the data. No data will be perfectly clean and consistent, but users will need to review any notes that may affect their work with the dataset. Therefore, throughout the process of reviewing, cleaning, and finalizing the data, build documentation with all relevant information for data users. This may include, but is not limited to, the following examples:

* Counts of all groups in the dataset: household, women 15–49, primary decisionmakers, children under 5, children under 2, all groups of farmers (total and per value chain commodity)
* Skip errors or other errors in programming discovered during or after fieldwork
* Numbers of outstanding inconsistencies, such as age differences between the household roster and other modules, children whose weights were inconsistent with possible ranges for their ages, households that indicated no electricity but connection to the national grid, etc.
* Lists of added country-specific questions (as compared to the core Feed the Future ZOI Survey)

Analysts for the survey should also contribute to this document, including any relevant information discovered during analysis.