WFP

World Food Programme

SAVING LIVES
CHANGING LIVES

WFP Data Quality Guidance Note

For Food Security & Essential Needs Assessments and Monitoring

January 2025

Contents

Introduction

The WFP Data Quality Practical Guidance Note is an organisational tool that sets out to establish operational best practices for ensuring quality quantitative data in the Vulnerability Analysis and Mapping (VAM) and Monitoring units, and ultimately to ensure the validity and reliability of data-driven decision-making.

The Country Office survey teams (usually under VAM and Monitoring functions) must plan and implement systematic quality assurance procedures to prevent, identify, and correct errors and ensure best practices are being followed throughout the data collection cycle. Implementing effective and efficient data quality assurance strategies will lead to high-quality data and timely and accurate results.

Data quality assurance starts with file management and survey questionnaire design and continues throughout enumerator training, pilot testing, high frequency checks (HFCs) during data collection to flag irregular reported data points, data cleaning and the final data analysis. The data quality assurance process should also trigger a timely review of the questionnaire/programmed instrument and enumerator qualification with potential refresher training, and any other measures required to correct any inconsistency from the very beginning of the survey process. At the end of this process, a reproducibility package should be created, including a final raw and cleaned dataset, syntaxes including all cleaning procedures and explanatory notes, the final coded questionnaire used, the report and any relevant files.

DATA RELIABILITY refers to the consistency of the data and the extent to which it is free from error. When data are reliable, using the same test multiple times would result in the same outcome every time (reproducibility). Meanwhile, the validity of data measures how true the data measured are.

HIGH QUALITY DATA refers to data that is not systematically biased and accurately represents the population(s) of interest, with adequate coverage of all aspects related to the research questions or goals of assessment and monitoring activities.

The target audience of this guidance note is primarily WFP staff involved in food security household data collection or analysis, either in VAM or monitoring functions. This is because many of the tools and platforms mentioned herein apply to WFP staff (and data collection partners and contractors), such as the VAM Resource Centre, Survey Designer, Data Library, MoDa, and SharePoint. Some of these resources are either publicly available or access can be granted based on request; however, WFP staff are able to access these resources with their WFP credentials. In any case, much of the guidance is also applicable to others in similar roles at other humanitarian and development agencies that collect food security data.

How to use this guidance

This guidance is designed to be used in a phase-wise or modular approach. Depending on previous experience with household surveys, it may not be necessary for all analysts or users to review the full document from beginning to end. Instead, they should review the checklist in the next section, and if they need more information, they can click the links to the headers to read a more detailed description of the steps to follow.

It should be noted that the checklist has been designed to be short to encourage analysts to review the more detailed relevant sections to ensure that the latest points have been considered.

It is important to note that some concepts are repeated throughout the phases because 1) analysts may not have read the entire guidance, and 2) there are multiple steps when quality checks should be conducted. For example, setting constraints are mentioned in the section on questionnaire design, and it is also mentioned in the HFCs and data cleaning sections. This is since constraints are relevant across various stages of the data collection cycle.

The guidance note is structured of three key phases:

1. **BEFORE DATA COLLECTION**
2. **DURING DATA COLLECTION**
3. **AFTER DATA COLLECTION**

It is important to highlight that by the time the data collection is finished (Phase 3), it is very difficult to course-correct and fix specific types of errors in the data collected. Therefore, users are encouraged to review each section of this document at least once in the survey planning stage to ensure potential data quality issues are captured and addressed during questionnaire design, enumerator training and high frequency check processes, leading to only little work during the data cleaning stage.

In addition, it is important to note that this document does not go into some of the practical aspects of planning and conducting assessments, such as budgeting, sampling and weighting, procurement of service providers for data collection, nor the planning, coordination and timing of assessments (e.g., considering seasonality). Neither should this guidance be used to design the data collection tool or training, as it only includes elements related to data quality. Instead, refer to other key resources on the VAM Resource Centre, especially those on the planning for Emergency Food Security Assessments (EFSA) and the Comprehensive Food Security and Vulnerability Assessments (CFSVA). For more information, please refer to the Planning Phase - WFP VAM Resource Centre.

Bon

Checklist

BEFORE DATA COLLECTION

1. FILE MANAGEMENT

- ☐ Have you created **dedicated shared folders** on an agreed-upon platform for saving all relevant documents with **appropriate access permissions**?
- ☐ Have you added the survey as planned in Data Library (guidance here)?

2. QUESTIONNAIRE DESIGN AND PROGRAMMING

- ☐ Have you created/updated the **assessment analysis plan** that addresses the study objectives?
- ☐ Have you ensured only questions relevant to the **study objective** are included?
- ☐ Have you used the standardized modules and variable names as per **Survey Designer** and the **Codebook**?
- ☐ Have you ensured that you included **informed consent** to participate in the survey?
- ☐ Have you ensured that there is a **uniquely identifiable survey ID number**?
- ☐ Have you reviewed and **minimized open-ended questions** and responses?
- ☐ Have you checked that all key questions are standardized and set as **mandatory** to avoid unexpected missing values?
- ☐ Have you ensured that there are **no umbrella (yes/no) questions** allowing enumerators to skip key modules?
- ☐ Have you considered the **order** of the modules based on logic, context and sensitivity?
- ☐ Have you added and tested the **skip logic, constraints, and warning messages**?
- ☐ Have you **contextualised** modules with use from thematic experts, where relevant?
- ☐ Have you finished the **translation** in all languages, where relevant?
- ☐ Have you **tested the programmed tool** on an electronic device?
- ☐ Have you tested the flow and length of the questionnaire and only included the **necessary questions**?

3. ENUMERATOR TRAINING

- ☐ Have you considered **context-appropriate backgrounds** during the recruitment of enumerators, including gender, ethnicity, language?
- ☐ If collecting anthropometric nutrition data, have you included trainers that are SMART experts and ensured you have a sufficient number of MUAC tapes for training and data collection?
- ☐ Have you prepared/updated the **training materials** (e.g., manual, agenda, presentations) in the necessary languages?
- ☐ Have you prepared a document with contextualised examples of **small quantities** for the training of the food consumption score module?
- ☐ Have you discussed **typical consumption habits** for the context?
- ☐ Have you discussed the **correlation** between the key food security modules?
- ☐ Have you discussed the role of enumerator **observation**?
- ☐ Have you discussed the importance of **probing**, especially when inconsistent responses are provided?
- ☐ Have you reviewed the response options and the use of "**other, please specify**", and "not applicable?
- ☐ Have you reviewed the use of **enumerator field notes** and reporting?
- ☐ Have you included **practical exercises and role-playing scenarios** in the training?
- ☐ Have you included a **post-training test** with focus on testing all enumerators' understanding of the tool, sampling and key modules, ensuring only enumerators who passed the test is sent for data collection?
- ☐ Have you **finalised the questionnaire** incorporating any potential inputs coming from the training and pilot testing, as well as feedback from the enumerators?
- ☐ Do you have an **enumerator management and communication plan** in place, including a **feedback mechanism** for enumerators to report issues during data collection?

DURING DATA COLLECTION

1. HIGH FREQUENCY CHECKS

- □ Have you checked the **number of completed interviews** against the designed sample?
- □ Have you checked the distribution of key variables and identified **erroneous values and outliers?**
- □ Have you checked for **missing values in key variables?**
- □ Have you checked the use of **special values, such as “don’t know” and “other, specify”?**
- □ Have you checked for **duplicate records?**
- □ Have you checked the **survey duration** and flagged surveys that are significantly shorter or longer than average?
- □ Have you reviewed **anomalous consumption patterns?**
- □ Have you reviewed **anomalous coping strategy behaviours?**
- □ Have you checked that the **expenditure module responses make economic sense?**
- □ Have you checked for **illogical responses by triangulating the data?**
- □ Have you checked **enumerator performance**, including daily completion, missing data, response inconsistencies, and flagged errors?
- □ Have you provided **regular feedback** to team leaders about enumerators based on their performance checks?

2. ISSUE LOG

- □ Have you documented all potential errors in an **issue log?**
- □ Have you categorized issues by type (e.g., data quality, technical problems) for easier resolution?
- □ Have you **communicated all issues** to the field supervisors for clarification and correction as planned?
- □ Have you established a **timeline for issue resolution** and follow-up to ensure corrections are made promptly?

AFTER DATA COLLECTION

1. DATA CLEANING

- ☐ Have you completed the designed **sample**?
- ☐ Have you **documented any deviation from the designed sample** and the potential impact on the analysis?
- ☐ Have you reviewed and addressed **special values?**
- ☐ Have you dealt with **anomalous expenditure values?**
- ☐ Have you verified that data are consistent through **triangulation?**
- ☐ Have you **documented key steps in your syntax and lessons learned** to be reviewed for future assessments?
- ☐ Are you keeping a **data cleaning log** that includes steps taken and rationale for decisions made during the cleaning process?

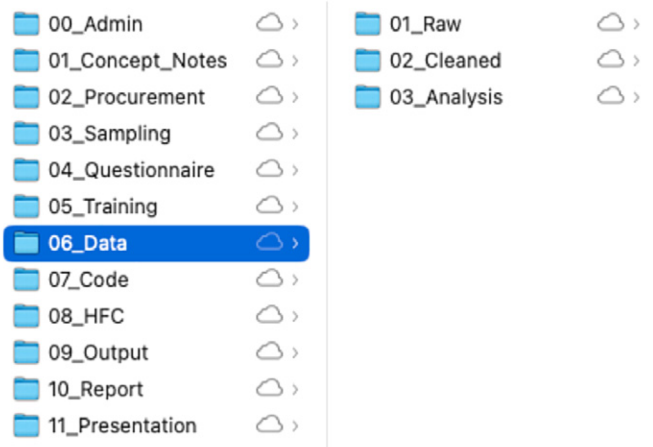
2. DATA AND DOCUMENT MANAGEMENT

- ☐ Have you uploaded the raw and cleaned datasets, and scripts processing the data to **dedicated folders?**
- ☐ Have you uploaded the final versions of the raw and cleaned datasets, scripts, questionnaire, and other related documents to **DataLib?**

Before Data Collection

FILE MANAGEMENT

Folder Organisation



00_Admin
01_Concept_Notes
02_Procurement
03_Sampling
04_Questionnaire
05_Training
06_Data
07_Code
08_HFC
09_Output
10_Report
11_Presentation

01_Raw
02_Cleaned
03_Analysis

Data quality starts with proper data management, keeping key files such as the terms of reference, the finalised questionnaire, raw and cleaned datasets, and syntaxes in dedicated and separate but organised folders. For WFP, this is likely to be using a combination of SharePoint and Data Library but may differ depending on the Country Office.

- **Version Control:** There needs to be a system in place for version control, using dates and indicating clearly which documents are the final. It is recommended to make one folder for each assessment using a shared space, e.g., calling it "CFSVA 2024" or "2024_CFSVA", which should contain all documents relevant to the assessment. It should include sub-folders for administrative and procurement, methodology, sampling strategy, raw and cleaned data, syntaxes used, results/output file(s) and report produced etc.

- **Document management:** During the assessment, the folder should contain working documents in an organised manner. Each quality check on newly downloaded raw datasets should be labelled based on the reference date and saved in the appropriate folder (e.g., 'Day 3' or '20240203'). As new data arrives, save these files in separate subfolders labelled by date (e.g., 'Day 4' or '20240204'), along with updated checks and outputs.

- **Archiving:** After the exercise is done, archive all files that are no longer relevant and ensure that you mark documents as final.

File management is crucial for preventing data loss, ensuring data integrity, and creating a 'reproducibility package' that allows colleagues to understand and replicate processes, and conduct additional analysis if needed, particularly in a work environment where colleagues are often moving between Country Offices.

Data Library

At this preliminary stage, the country office should create a space for the survey in Data Library marking it as "planned", and add any information available at the time. It is important to do this preliminary organisational step as creating a folder in Data Library will ensure storage in a secure place, and archives WFP resources in one repository. For more information, see the Data Library Guidance.

Note that Data Library is a WFP data repository, and WFP accounts are already part of the registered domain; however, external users can also be granted access by creating an account and will, by default, only be able to see public resources and meta-data. This is also the platform used if the Country Office wishes to share data with external stakeholders, e.g. donors, IPC team etc. After the data collection phase has been completed, the final raw and cleaned datasets, along with questionnaires, scripts, and codebook should be uploaded to Data Library.

PRELIMINARY STAGE

- **Create a Survey Space:** Early in the process, create a space for the survey in Data Library, marking it as "planned" and adding any available information. This helps in organizing the survey from the outset and keeps all relevant documents in one place.

- **Benefits:** This preliminary step ensures secure storage of documents, and archives WFP resources in a centralized repository.

- **Guidance:** Refer to the Data Library Guidance for detailed instructions on setting up and managing your survey space.

FINAL STAGE

- **Upload Final Documents:** After data collection, upload the final versions of raw and cleaned datasets, questionnaires, scripts, and codebooks to Data Library. This ensures that all crucial documents are stored securely and are easily accessible for future reference.

- **Access Control:** Ensure that data privacy guidelines are followed by restricting access to these files to authorized personnel only. Note that while WFP accounts are part of the registered domain, external users may be granted access to public resources and metadata by creating an account.

Following these steps will ensure that all survey-related documents are securely stored, easily accessible, and well-organised, supporting efficient data management and future reproducibility.

QUESTIONNAIRE DESIGN

WFP Survey Designer is the starting point for questionnaire design, where the main modules and survey elements can be directly downloaded and customised to the country context while maintaining the standard module questions.

The Survey Designer then produces two possible outputs: a Word Document and/or an XLSForm in four possible languages (English, French, Spanish and Arabic). While the XLSForm is the minimum output to be able to be uploaded to MoDa for administration in the field, some Country Offices may also choose to maintain a copy of the questionnaire in Word format as well as this is easier to read, e.g. for enumerators.

Note that if you are using Survey Designer, the following checks, as well as many indications related to skip logic, constraints and warnings, should not be needed as they are already included in the standard modules in Survey Designer. In general, it is strongly advised to follow the standard modules and response options in Survey Designer. Additional context-specific modules that are not in the Survey Designer can also be added using the same format and the standard variable names from the codebook if possible.

General

❑ **Use WFP standard modules and codebook.** Please check WFP Survey Designer and Codebook for the latest modules and programming of XLSForms, which can then be used in MoDA. Whether you are using survey designer, or you have a non-standard module/survey, make sure to check that:

- The response options should be consistent throughout the questionnaire. This includes standardising the response **option values**, e.g., 888 = Do not know, and 999 = Other, please specify.

REMINDER: While it is usually recommended to use the same survey modules and questions as used in previous assessment and monitoring activities, a periodical review is necessary to check the relevance and adjust modules according to latest guidance and needs.

It is strongly advised to develop an **Analysis Plan** prior to designing any data collection tools. This will guide the key analytical questions and help to answer which information is needed based on the objectives of the assessment. This can also be used to shorten existing tools by deleting questions that are not needed.

- When there are **special values** like “other, please specify” answer options, ensure that there is an open-ended text question following.
- **Filter questions:** In checking that the modules are up-to-date and correct, please ensure that there are no “filter” yes/no questions before key food security modules, such as the reduced food-based coping strategies index (rCSI) and Livelihood Coping Strategies index (LCS) asking, *“In the past 7/30 days, were there times when you did not have enough food or money to buy food?”* Introducing this filter “yes/no” question before the main module enables enumerators to skip the entire module, and this **should not be permitted under any circumstances**. All the standard questions for these key food security indicators must always be asked in full.
- **Include basic survey elements**, some of which are normal fields such as survey introduction, unique survey ID, enumerator ID, and GPS coordinates if allowed in the context, while others are metadata (e.g., start and end time, today) etc.

❑ **Obtain informed consent.** No further questions should be asked of the respondent if they do not give their consent to participate in the survey. There is no point in collecting data that cannot be used, and it would be unethical to do so. Thus, ensure that consent is marked as mandatory and that enumerators are properly trained to administer it at the start of every survey.

❑ **Apply eligibility filter questions if applicable.** If the survey has inclusion criteria, ensure that relevant questions are moved to the beginning of the tool and a skip filter is added. E.g. if the survey is only for displaced populations, the survey should be ended if the respondent does not match this criterion to avoid inclusion errors.

❑ **Consider the order of the modules.** Sensitive questions should be placed at the end of the questionnaire, such as the Household Hunger Scale (HHS). This may also include questions on protection, safety, and social cohesion. Equally, it is recommended that the modules for core WFP indicators be towards the beginning of the questionnaire to avoid respondent fatigue at the time when the key areas of interest to the survey are being collected.

❑ **Include only questions that are needed** and intended to be analysed and reported on, including all mandatory indicators and key socio-demographic indicators.

- For **WFP food security assessments** these include:
  - → The Food Consumption Score (FCS)
  - → The reduced Coping Strategies Index (rCSI)
  - → Livelihood Coping Strategies – Food Security (LCS-FS)
  - → Food Expenditure Share (FES)
- For **WFP essential needs assessments** these include:
  - → The Food Consumption Score (FCS)
  - → The reduced Coping Strategies Index (rCSI)
  - → Livelihood Coping Strategies – Essential Needs (LCS-EN)
  - → Economic Capacity to Meet Essential Needs (ECMEN)
- For **WFP monitoring activities** (site visits, baseline, endline and follow-ups) please see the Monitoring Phase, as well as the latest WFP Indicator Compendium. Base content in Survey Designer provides standard topics for inclusion related to Process Monitoring (e.g. Distribution Conditions) considered to be minimum requirements.

❑ **Minimize open-ended questions and responses.** Try to avoid the use of 'other, please specify,' 'don't know,' and 'refused to respond'. For regular assessment and monitoring questions, respondents should be aware of most information related to their households. The inclusion when not relevant could lead to issues getting meaningful analysis from the indicator. In addition, remember that the 'other, specify' answers imply extra work: they should be reviewed and recoded, if necessary, in the cleaning stage, before analysing and reporting on the main questions. While some of this can be mitigated during the enumerator training, it should be considered also during the questionnaire design.

❑ **Check that all key questions are set as mandatory (required) when programming the XLSForm to avoid unexpected missing values.** In most cases, it is recommended to set all questions as mandatory to avoid data gaps; at minimum, the core indicators' modules (e.g., FCS, rCSI, LCS, and expenditures) must be programmed as mandatory, so that it is not possible to skip any of the modules. However, there are some exceptions to the rule, for example, setting sensitive questions as mandatory can negatively affect the response rate and can cause issues in the communities surveyed. Also, setting the GPS coordinates as mandatory in areas where this is not sensitive can still create issues if there are areas with limited connection.

❑ **Logical order of key modules.** If FCS and HDDS are both included, it is recommended to collect HDDS right after FCS where the respondent will easily remember the food items in each group and for the enumerator to use the FCS responses for easy consistency checks and probing during the HDDS data collection.

- ❑ **Ensure contextualisation.** Certain questions/modules/indicators need to be contextualised to the country context to get quality data. These include, for example:
  - **Field team information:** team lead and enumerator names after the enumerator training when teams have been decided.
  - **FCS:** examples of food items per standard food group. While this should mainly be done during the tool design stage, ensure to revise the coded tool if needed after the enumerator training.
  - **LCS:** decide on coping strategies to be collected and the contextualised severity weights[1]. This should be done when designing the tool.
  - **Expenditures:** add examples of food and non-food items. If applicable, ensure that the currency is explicitly mentioned in the questions, especially in contexts where the currency changes or where different currencies are accepted.
  - **Perceived needs:** select core areas of relevance
  - **Education:** ensure that education levels reflect the country's context.

WARNINGS VS. CONSTRAINTS:
Constraints are used during data collection as "hard" prevention measures, whereas warnings are more "soft" measures that can flag potential errors to enumerators, and remind them to check the consistency of answers, as well as their understanding and abilities through a check-clarify-correct system. In this way, warnings can work as a corrective measure to help enumerators by making them aware if the answers given by the respondent lack logic.

It is recommended not to use too many constraints as it may force data responses in one module where the actual mistake could be from a previous module. Also, it can also mask actual enumerator performance for the Data Analyst.

Questionnaire Programming

Logic-based techniques (e.g., validation checks and skip patterns) among modules cannot always be standard as assessment and monitoring questionnaires differ from one country to the next, depending on information needs that extend beyond the standard corporate modules. However, constraints should be set prior to the start of the data collection exercise and adjusted during the testing and piloting phases (noting that questionnaire testing should take place in two phases: prior to the enumerator training and again during the enumerator training).

Skip logic: Ensure that skip logic is in place where relevant. If it is designed to skip an entire module under a specific condition, make sure all relevant questions are grouped and skipped at the same time. In Survey Designer, skip logic is already in place for standard module.

Constraints: Ensure that constraints are in place and set up carefully. It is, however, a trade-off on how many constraints should be implemented as too many constraints may mask which enumerators have a good understanding of the module and which enumerators need additional training. Thus, some Analysts may prefer to not apply constraints, or instead to apply warnings (see box to the right) preferably on indicators where inconsistencies are easily detectable, and instead use the high frequency checks to identify which enumerators need to be retrained. This is ultimately up to the Data Analyst to decide. Constraints are indicated with the ⊗ bullets.

[1] Note that the country context can change over time, e.g., if conflict breaks out/escalates or another shock occurs. Livelihood-based coping strategies, in this case, must be reviewed to make sure they are still capturing the strategies used by the population surveyed. If new strategies are used that are not in Survey Designer, contact HQ to approve that the new strategy can be considered as a livelihood coping strategy and to assign a standard code.

For the latest recommended constraints, please download the module directly from Survey Designer. Further guidance on XLSForm programming can also be found here: XLSForm Docs.

WARNINGS: In addition, it is important to set warnings for illogical values, and to flag to enumerators to double-check the illogical responses before proceeding. Note that the mistake can be linked to previous answers, e.g., if a respondent says that no cereal was consumed yesterday in the HDDS module, but this is flagged because the answer for cereals consumption in the FCS module was 7 days, then the enumerator should probe to know which response is correct and go back to correct any previously misreported information, if necessary. Warnings can be included throughout the questionnaire, **see the below examples flagged in red.**

KEY DEFINITIONS: It is vital to data quality that all enumerators are aware of the key definitions of a household survey, since this is the basis of all the data collected. These are:

- **HOUSEHOLD:** A household is made up of one or more individuals (persons) living in the same dwelling (people living under the same roof), typically sharing the same meals and under the responsibility of one person (head of household); household members may or may not be related.

- **HEAD OF HOUSEHOLD:** the head of household is the individual that typically has the responsibility of decision-making in the household and must be part of the household.

Examples of relevant constraints and warnings

DEMOGRAPHIC SECTION

- **Age:** to be constrained to between 0-99.
  - If the respondent is under 18 years of age, flag to double-check the age. In most countries, for protection reasons, respondents should be over the age of 18 (or sometimes 15 or 16, depending on the country's context). That said, in extreme cases, children (such as unaccompanied minors) can be interviewed to ensure their extreme situation is not overlooked.

- **Head of household:** ensure that only one head of household is indicated.

- **Total household size:** the total household size should be automatically calculated after asking the number of household members in each age group. Afterwards, the total number of household members should be asked as a separate question to validate that the indicator is correct. Household size should always include the household head in the count.

- **Children:** the number of children not attending school cannot be greater than the number of children in the household.

- **Disability:** the number of household members with disability/chronic illness cannot exceed the number of household members.

FOOD SECURITY INDICATORS

Food consumption: values for the days should be constrained to a range between 0-7 days (inclusive). 0s across the module (total FCS of 0), not possible unless extreme famine conditions.

- 7s across the module (total FCS of 112), unlikely that households ate every food group in the past 7 days.
- Very low consumption of staples (4 days or less) (see box to right).
- Very low consumption of sugar and oil (3 days or less) in countries using the high FCS threshold.

FOOD SOURCES

- It is recommended to code the XLSForm to only allow the hunting/fishing option of food sources for the protein food group.
- If the food consumption frequency is 0 for a food group, then the food source should automatically be "Not applicable," and this question should not be asked.
- **FCS-N:** values should be constrained to ensure that the figures reported for the subgroups do not exceed the number of days reported for the main food group.

HDDS

- Add warning if the respondent reports that the household consumed nothing yesterday (0) and if they consumed all 12 items yesterday (12).
- Add warning when a household reported 7 days of consumption of a food group in the FCS module but did not consume the same food group within the last 24 hours (HDDS).
- Add warning when a household reported no consumption of a food group in the last 7 days (FCS) but reported consuming that food group within the last 24 hours (HDDS).

ANOMALOUS CONSUMPTION: Consider adding a warning message to the XLSForm to ask to verify anomalous consumption levels, such as very low consumption of staples e.g., by asking "Why were cereals consumed less than 4 days over the past 7 days? What did the household eat instead?" In most countries, households consume staples daily, so consumption of less than 4 days should be flagged as potential underreporting. However, in some contexts pulses could replace cereals consumption in some/all days.

Food consumption in HDDS vs. FCS: Note that while the HDDS data can be validated using FCS data, the opposite is not possible since HDDS considers food consumed in small quantities and anyone in the household (not the majority like FCS). For more elaboration, refer to the HDDS guidance (VAM Resource Centre).

rCSI: values for the days should be constrained to a range between 0-7 days (inclusive).

- Add warning when a household has no child under 5, and the strategy on 'restricting consumption by adults in order for small children to eat' is > 0.

LCS

Children:

- If the household does not have children, the child-related strategies (ChildWork, ChildMarriage, etc. should be marked as N/A); set a warning to flag this for review when responses other than N/A are provided.
- If a household has no school-aged children, withdrawing children from school as a coping strategy should be not applicable; set a warning to flag this for review when responses other than N/A are provided.

Not applicable:

- If a household responds that 3 or more of the 10 strategies are N/A, set a warning to flag to review the responses and double-check that they have understood the question and the response is true "not applicable". Note that if a high usage of not applicable is true, it may warrant a need to review and change of the strategies used to capture livelihood coping.

Expenditures:

- 0s across the module for food expenditures (total food expenditure is 0); it is highly unlikely that a household would have spent nothing on food.
- 0s across the module for non-food expenditures (total non-food expenditure is 0), which is similarly unlikely.
- Expenditures with values lower than the lowest unit of currency.
- Very high values for individual expenditure items, for both food and non-food. It is highly recommended to add actual, contextualized warning thresholds as expenditure data is among the most difficult type of data to clean.

Translation

The questionnaires must be translated into relevant local languages directly in the XLSForm to maintain the precise meaning of the questions and ensure all enumerators are asking each question in the correct way. In this process, it is recommended that one staff member translates the questionnaire and then another staff translates it back into the original language as it minimizes the risk of translation errors. Note that in surveys using multiple languages, it is important to ensure that changes to the questionnaire are copied across all languages.

Beware that issues arise if questions are not correctly translated. E.g., "selling the last female animal" as a livelihood coping strategy is asked as **"selling female animals"** in the survey language will give a completely different meaning and the question will no longer reflect emergency coping. This is not possible to clean afterwards and will mean that the data for this question needs to be discarded.

Sometimes, it is not feasible/practical to translate the questionnaire into every local language employed in a country. Instead, it is common practice to have the questionnaire standardized in the national language.

BEST PRACTICES FOR TRANSLATION:

The golden standard is to have the questionnaire available in all relevant local languages. However, the minimum recommendation is to have the questionnaire in main national/local language of the relevant populations.

DATA COLLECTION TOOL AND TESTING

Due to the length and comprehensiveness of standard WFP questionnaires, it is **always recommended to collect data using tablets** rather than phones. This since the screen, keyboard etc. is more suitable for data collection. Using mobile phones with small screens comes with data quality risks such as typos due to a smaller keyboard, higher likelihood of enumerators not following logic/consistency as they can only see very little sections of the tool at the same time, enumerator fatigue leading to non-correction of erroneous values (if having to click hundreds of times to correct a previous value) etc.

The programmed questionnaire should be bench-tested after it is uploaded to MoDa and deployed without errors to make sure all the questions and choice options are programmed correctly, reflecting the designed validation constraints and skip patterns. During this phase, check that umbrella questions are not allowing enumerators to skip entire key modules. The tool should be deployed to an electronic device and tested from start to finish by multiple team members. It is essential to do this before starting an enumerator training.

Additionally, it recommended testing the questionnaire using different modules and versions of electronic devices to confirm that the questionnaire functions correctly in different environments and under varying conditions to ensure reliability during actual data collection and to test the duration of the interview.

ENUMERATOR TRAINING

It is important to stress that the enumerators are the primary point of control for data quality assurance once the data collection has started, and that sufficient number of days, relevant training capacity etc. should be allocated to the training. Furthermore, to ensure that only qualified enumerators are selected for field data collection, **Country Offices must train more people than the required number of enumerators.** At the end of the training, all enumerators must pass a final test in the same language as the data collection will take place before being sent to the field. Sending enumerators to the field that are not able to pass the test will compromise the data quality.

Overall

- **Introduction:** The training should start with a brief presentation on WFP and its key principles, as well as the objective of the survey and how the data will be used in order to motivate the enumerators and to stress the importance of collecting high-quality data.

- **Sampling:** During the training, it is important to train the enumerators on the chosen sampling methodology, highlighting that if the sampling strategy is not followed it will have major consequences for the data quality as the survey results will no longer be representative. **Correcting wrong sampling at a later stage is not possible**, and consequently the whole survey would need to be redone.

- **Use of 'other, please specify':** During the enumerator training, ensure that it is stressed to only use the "Other, please specify" option when necessary and to carefully check first that the response is not part of the pre-existing list. Enumerators are encouraged to use this option only when they are unclear about the categorisation of existing option lists, with adequate information after probing for the field supervisors to make informed decisions later. Please also ensure that enumerators understand well each item in existing option lists in modules where "other, please specify" is used. This is particularly important in modules that are used for vulnerability assessments or targeting and prioritisation, for example, income sources and the housing and WASH modules.

- **Observations:** Enumerators should also be encouraged to use common sense and observation when conducting face-to-face interviews and in probing about visual cues, such as household members, possession of assets, housing types, etc.

- **High frequency checks:** Once the data has been collected, it is very difficult to clean up potential errors. It should be stressed that consequently all interviews will be checked using high frequency checks including follow-ups and retraining with enumerators making frequent errors.

- **Final tool and practical exercises:** The final programmed questionnaire is a mandatory tool to be presented and piloted during the enumerator training. Note that enumerator trainings should always include practical exercises such as testing data collection and upload using the tablets.

- **Team leader responsibilities:** A specific session should be held for team leaders once selected (after the test) to ensure they are aware of their specific responsibilities including sampling, following up on high frequency checks etc.

Key Indicators/Modules

The following sections present key areas for the enumerators to pay careful attention to and probe respondents on while data collection is ongoing. Please visit the VAM resource centre to see the latest resources, including training manuals and PowerPoint presentations for each of the food security and essential need indicators that can be used for the enumerator training. The training presentations are available in English, French, Spanish and Arabic.

Food Consumption Score module: During the enumerator training, allocate time to discuss **consumption habits and what is considered typical/atypical** by area and population group (e.g., ethnicity, religion, etc.) to enable enumerators to probe respondents properly regarding food consumption and spending:

- Thoroughly discuss small quantities to ensure that only relevant consumption is considered in the module. It is recommended to give the enumerators a printed, contextualized version of the small quantities overview to bring to the field. In any doubt about whether specific, contextual consumption falls under 'small quantities', contact VAM HQ.
- In case of low cereals/tubers consumption, enumerators need to probe to make sure that the household did not consume staple foods every day, and if not, understand what they consumed instead. For example, it may be possible that the household consumed pulses/legumes (e.g., lentils or beans), in place of staples, which could be understood and accepted.
- For protein consumption (meat, fish and eggs), make sure that egg consumption is considered also in the main group. Very high consumption for seemingly poor households should be confirmed with the household to check that they did not include small quantities and only included foods consumed by the majority of the household.

Expenditure module: Discuss spending habits, costs of food and possible amounts to spend on certain food items as well as non-food items, in order to determine what is considered typical/atypical for the context:

- This needs to be discussed for the lowest possible expenditure, and the highest possible expenditure, at both ends, for the poorest and the richest. Remember to consider household size as larger households will likely have higher variable expenses.
- Meanwhile, discuss amounts that do not make sense, such as food expenditures of 1 or 2 when a unit of bread costs 50 local currency, noting that no expenditures should be lower than the lower unit of currency.
- Remind enumerators to take care in contexts where there are multiple currencies (border areas), or denominations (ex., Iranian rial and toman), to ensure that there is no misreporting due to different currencies being used in the survey responses.

Livelihood Coping Strategies module: During the enumerator training, the meaning of the **response options for LCS** needs to be elaborated and stressed, including the use of N/A:

- For certain strategies, like begging, or theft, it is rarely possible to exhaust these strategies (unless the household is located in extremely isolated areas where there is no one else to beg/steal from, or where the community reaches complete collapse and there is no one in the community to beg/steal from).
- In addition, N/A is not an option, as no matter how 'socially unacceptable' these coping strategies may be, they are still options. For these strategies, the answer should either be 'Yes' or 'No, we did not need to (or engaged in other strategies)'.

Data Consistency

While it may be difficult for enumerators to immediately spot these linkages, enumerators should be trained to notice and flag inconsistencies reported by the respondent, and probe further to get quality household data. It should be noted that issues can stem from other prior questions and the enumerator should probe to determine the correct information, and go back and correct any previously misreported information, rather than assume the correct answer.

The following are examples of inconsistencies to look for, however, they will depend on the actual questions selected. Country Offices are encouraged to pick and choose data consistency discussions as relevant to their tool/context:

- **Inconsistencies in the household demographics:**
  - Head of Household (HHH) with a marital status of 'single' cannot have household members marked as spouse/partner.
  - The father of the head of household (HHH) should not be younger than the HHH.
  - It is not possible to have more than one head of household.
  - If the head of household is a minor but the household has other adult members, confirm that the minor is indeed considered the head (i.e. responsible for all decision-making in the household).

BEST PRACTICES FOR DATA CONSISTENCY (WARNINGS VS. CONSTRAINTS): For example, if a respondent first says no children are in the household, but later says coping strategies related to children were applied, the enumerator should never assume which answer is incorrect but must always probe and let the respondent correct. Note that warnings are recommended rather than constraints for these examples since we cannot know which indication is the reason for the inconsistency – there could be children in the household not reported in the demographic module due to lack of clarity on the definition of a household, or there could have been children in the household during the last 12 months that are no longer living there. Instead, a warning message can be used to remind the enumerator of the inconsistency, and the enumerator can then use probing to correct the right indicator.

- **Inconsistencies between the household demographics and the coping strategies:** If a household reports not having any children, then child-related coping strategies, such as reducing adult consumption for the sake of children (rCSI), withdrawing children from school, moving children to a less expensive school, child marriage and child labour (LCS) should be marked 0 in rCSI and not applicable (N/A) in LCS. However, it is possible that child members mentioned in LCS strategies may not have been accounted for in the household roster if they have already left through migration, marriage, or being sent to live elsewhere.

- **Inconsistencies between food sources and the location:** Check the FCS sources thoroughly for illogical answers by using knowledge of the local context and triangulating with the other relevant data points gathered in the same interview.
  - For example, reporting of own production or fishing/hunting in deserts, urban or landlocked areas.
  - Certain food sources only make sense for some food groups, e.g., hunting/fishing only applies to 'meat, fish and eggs.'

- **Inconsistencies in the main source of food and assistance.** For households that received assistance and reported this as a food source, this data should also be consistent with the "Assistance" module. Also, only food groups that are given as assistance should have assistance as main source of food, e.g. if no agencies are providing fruits as assistance, this should not be chosen as main source of fruit.

- **Inconsistencies between the food consumption and food expenditures.** The respondent reports not having spent any money on perishable food products (e.g., dairy products, meat, fruit and vegetables) in the past 7 days in the Expenditure module but reports frequent dairy consumption with cash purchases being the main source in the consumption module, or vice versa – reports no food consumption while having large expenditures.

- **Inconsistencies between assets and asset depletion in the livelihood coping module.** If a household reports current ownership of assets such as household items electrical items, or transport assets, then the related livelihood coping strategies on domestic assets and productive assets, respectively, cannot be recorded as exhausted or not applicable. Similarly, if a household reports reliance on rural-related coping strategies, such as selling animals there should be some consistency in the ownership of livestock/agricultural assets.

- **Inconsistencies between housing, assets and livelihood coping.** The type of household reported should align somewhat with the reported ownership of assets and livelihood coping strategies. For example, a household that reports living on the street or having depleted nearly all assets due to food insecurity, may not report the ownership of many assets.

- **Inconsistencies between expenditures and other modules, e.g.,**
  - Amount spent on 'rent' sounds reasonable considering the housing type and residence area.
  - Amount spent on 'electricity' sounds reasonable considering the housing type and household assets ownership.
  - Amount spent on 'education' sounds reasonable considering the school type, and the number of school-aged children or adults undergoing education.

- Inconsistencies between household hunger, and other food security indicators. Households reporting very high HHS would be expected to also report severe food insecurity across other food security outcomes including low food consumption, and high levels of coping.

- Inconsistencies between food consumption and dietary diversity. If a household reports having consumed an FCS food group all 7 days (e.g. oil), then the corresponding food group in HDDS must be counted as yes.
  - However, it should be noted that small quantities consumed by anyone in the household are counted in HDDS, so it is not possible to apply the reverse logic; meaning that HDDS can have a value of 1, while the corresponding food group in FCS can still be 0.

Enumerator's notes: It is recommended that enumerators and field supervisors take notes during the data collection about unclear questions, incomplete response categories, or awkward question flows, starting from the piloting during the enumerator training. They should also record qualitative information to explain the quantitative data. An easy way to administer this information is to add a text variable at the end of the coded tool where enumerator comments can be added. Enumerators' notes should be presented and discussed during daily debriefing sessions led by the team supervisor. Decisions should be properly documented and shared within the team.

Questionnaire Finalisation

During the enumerator training, enumerators should practice administering the questionnaire using tablets in the local language(s), and should flag any issues with skip logic, warnings, constraints,, translation, etc. Any issues flagged and additional updates agreed upon during the training (including context-specific warnings for Expenditure module, suggestions for local food items etc.) should be addressed in the finalised questionnaire. The programmed questionnaire should then be uploaded to MoDa again, replacing the old version(s) and ensuring that the enumerators have downloaded the correct and most updated version to the tablets for use in the field. It should be the responsibility of the Team Leader to verify that all enumerators in the team have a fully functional tablet with the correct form before going to the field.

Pilot Testing of the Final Tool

While pilot testing "in the field" with sensitised households near the training centre is encouraged as part of the training and tool testing/finalization, this may not always be possible to effectuate. If not, it is recommended to test the final questionnaire in the training venue. During this session, it is recommended to agree on a minimum acceptable time needed to fill one form, which should be used as a threshold during the high frequency checks. For **remote surveys**, pilot testing should always be done with the actual phone number using the actual tool; at least 25 phone calls per enumerator are required to test the performance.

Post-Training Enumerator Test

At the end of the enumerator training, it is recommended to test the knowledge of each enumerator to ensure that they gained the necessary knowledge before being sent to the field. The test should be prepared prior to the training and can be adapted in the final days of the training, after understanding the main issues that need to be addressed. It is vital that the test is contextualised to the tool, and that at least 1-2 questions on each key food security indicator are included. The test should avoid theoretical questions and questions that do not measure the enumerator's capacity to administer the questionnaire properly, e.g., questions on quantities "How many questions are there on coping?". It should instead focus on key definitions, practices, and lookouts from each of the modules. Depending on the size of the enumerator training cohort, it may be easiest to administer if coded in an XLSForm before the end of training and including automatic grading to make it less work-intensive.

Only enumerators that pass the test should be sent to the field to collect data or data quality will be compromised. If many enumerators have mistakes in the same module, consider re-training that session.

Note that it is often recommended to **assign the best performing people as team leaders** due to their importance in the data collection in terms of ensuring that the sampling method is followed, conduct follow ups after the high frequency checks etc.

Dropout and Replacement Roster

No enumerator training will lead to 100% of enumerators passing the test. Thus, it is advised to invite an additional margin of potential enumerators to the training; there is no fixed percentage for this, and it depends on context, budget, etc.

Further, it is recommended to account for enumerator dropout either during or after the training, especially in larger surveys or in cases where the survey has a tight deadline, e.g., finalisation before Ramadan or in time for a Humanitarian Needs Overview (HNO) or an Integrated Food Security Phase Classification (IPC) analysis.

Additionally, in instances where the team leaders and/or high frequency checks find that an enumerator is committing many data quality errors during the data collection which cannot be dealt with through additional training, it may be necessary to terminate the enumerator's contract/engagement.

For these reasons, it is recommended to train (and then test) a higher number of enumerators than needed, invite the best to collect data in the field while maintaining a roster of trained potential replacements that have passed the test.

Enumerator Management and Communication Plan

Before sending the enumerators to the field, it is essential to have a plan in place for enumerator management. This could take the form of a final list that includes essential information, such as enumerator names, enumerator ID, the area(s) each has been assigned to, which team they are in, who is the team leader, how to reach out for clarifications, and contact details. Team progress should be measured against the planned (usually found in the Assessment Terms of Reference), and can be monitored manually or virtually through a dashboard depending on connectivity.

In areas with limited connectivity and where data cannot always be uploaded daily, it is important to have a plan in place to ensure that teams are progressing as planned, e.g. the team leader may be responsible for reporting daily progress via Whatsapp.

Note that the high frequency checks require a way to communicate issues back to concerned enumerators. Therefore, it is essential to have planned for the best way of communicating with team leaders before sending the teams to the field. In this plan, it is important to assign specific people/roles to each step. For instance, if there is a language barrier, the Data Analyst may need to communicate to a national colleague assigned by the Country Office, who then communicates to the team leader. The latter step could be through debriefing calls to the team leaders within an arranged timeframe daily during the first week of data collection where feedback will be more frequent. In addition, before setting out to the field, it is useful to establish a clear timeline for error correction, so the field team is aware of potential timely actions that will be needed.

During Data Collection

CHECKS CONDUCTED BY TEAM LEADER

From the beginning of the data collection, the team leaders will work as the first step of data quality. The team leader will be responsible for ensuring that the sampling plan is being followed, which is key to guarantee representativeness of the data. Furthermore, the team leader should support each team member during the data collection. This includes answering technical questions, joining interviews with each team member to monitor their interviews, conduct spot checks, ensuring all interviews are done properly, following up on flags during the high frequency checks etc.

HIGH FREQUENCY CHECKS

High frequency checks (HFCs) are regular checks of the quality of the collected data, conducted in real-time during the data collection and communicated back to the field. **Note that setting up HFCs takes time, and it is recommended to start developing these at least 2 weeks prior to the expected start of data collection.** This includes developing a standard script or a dashboard with all checks done. Remember to reflect any potential last-minute changes to the tool, thresholds discussed during the enumerator training etc. in your final HFC script/dashboard. This includes suggested contextual thresholds for food consumption, expenditures etc.

Ideally, HFCs should be done daily from the start of data collection to detect and correct potential data quality issues as early as possible; **HFCs should be conducted by an experienced Data Analyst during the data collection with special emphasis on the first 2-3 weeks.**

FULL CYCLE DATA QUALITY APPROACH: Note that if the previous sections of this guidance have been applied properly during the first phase of the data collection (setting appropriate constraints, warnings, training and testing enumerators well etc.), fewer flags in the HFCs can be expected.

As HFCs are used to avoid repeated mistakes during the entirety of the data collection, these checks are most important during the first weeks of the data collection and the frequency can be reduced over time as data quality issues have been dealt with. During the end of the data collection, HFCs are no longer useful and only data cleaning can be used to correct errors.

In an ideal scenario with full network coverage, the process should be as follows:

- All data should be uploaded to the server by the end of each day's data collection for the HFCs to be done during the evening by the Data Analyst.
- After the review, the Data Analyst should prepare an overview highlighting issues for each enumerator.
- The overview should be shared with the respective team leaders in the field (or according to the plan established in a previous step). Depending on potential language barriers, this step may be done directly by the Data Analyst or by a national colleague assigned by the Country Office.
- The following morning should start with a debrief by the team leader, dealing with any data quality issues with the enumerators and the specific module(s) flagged, to avoid the same mistakes being repeated during the entire data collection
- The same process should be repeated daily, allowing real-time tracking of progress in the data collection, until no more issues are found.

In cases with poor network connectivity, enumerator teams should be encouraged to upload as often as possible. If possible, alternative options are recommended such as sending portable modems with the teams. If a survey team covers some areas with network coverage and some without, teams are encouraged to start in areas with coverage, allowing the HFCs to detect errors early in the process.

For **REMOTE SURVEYS**, besides conducting HFCs, audio recordings should be used to assess the performance of enumerators. Audio recordings should be checked at the beginning of data collection during the pilot testing and at random time points throughout the data collection. All issues identified and flagged should be logged in an issue log, to be considered in the next phase of data cleaning.

HFCs can take several forms: they could be done manually using any statistical software, or a dashboard could be created. The checks conducted should always be adapted to the tool used (based on the chosen modules and indicators) and the local context (e.g., realistic expenditure thresholds).

The aim of HFCs is to detect and flag potential issues from the outset of data collection to improve the data quality as early as possible. Thus, it should be highlighted that **HFC flags are not necessarily mistakes**, but data that warrants communication with the enumerator of concern to check if they are actual values or mistakes. E.g., if one enumerator is flagged for reporting very low FCS, it could be either because (s)he is misunderstanding the module, or it could be that (s)he is collecting data from an area that is extremely food insecure.

Consequently, rather than correcting previously collected data, which can lead to other data quality issues if based on guessing etc., **the purpose of the HFCs is mainly to correct future data collection following the discussion with the enumerator**. The exception is for data that are easy to correct, such as clear typos and mistakes done the previous day where the enumerator can still remember the correct value (e.g., 0 instead of 8, urban instead of rural, geographical information/p codes, information about household head size/sex/age etc.).

Data should be checked for differences by enumerator, team and by geographical area. The latter is particularly useful to detect training-related issues in cases where trainings were done in different geographical locations, and if there are natural geographical differences between areas.

With regards to the checks by enumerator, it is best to wait till there is a sufficient sample of interviews per enumerator after a few days, for example 10 interviews per enumerator, before making any judgment calls about the enumerator's understanding because the flag could be an actual outlier.

Generally, due to the high magnitude of incoming data and to avoid overburdening the Country Office and data collection teams, **the aim should be at detecting patterns** (e.g., one enumerator consistently reporting unrealistically low FCS, even if compared to the rest of the team), **rather than single outliers** (e.g., an enumerator has conducted one interview with a high number of N/A in the livelihood coping strategies module).

General Checks

❑ **Survey duration:** Check the survey duration for each of the interviews from start to end. For this, ideally, an estimate for an absolute minimum time needed to complete the survey should have been decided during the pilot of the final tool and communicated to the Data Analyst. After a few days of data collection, follow up with enumerators who are taking very little time or very long. *Note that long survey duration can be justified in areas with poor connectivity if submissions are not registered immediately. It can also differ depending on household size, meaning that it may take longer for larger households, especially in assessments where individual-level data (e.g. household roster, or nutrition outcome indicator, such as MUAC), are collected.*

❑ **Time of survey:** Check that surveys are done during normal working hours. The only exception is if data is collected on paper, and then entered during the evenings. In this case, the Data Analyst should be informed since this will impact the survey duration also.

❑ **Incomplete interviews:** A high number of incomplete interviews can indicate that something is wrong with the coding of the tool and should immediately be addressed.

❑ **Inclusion errors:** Inclusion of populations not falling under the survey (e.g., non-displaced in a survey for only displaced populations or wrong geographical area).

❑ **Number of interviews:**

- Examine the total number of surveys completed against the survey plan.
- Check the number of surveys by lowest sampling unit (e.g., cluster) covered and the gap between complete and planned.
- Check the number of interviews that each team and each enumerator is conducting daily to identify potential issues against the planned. This should be done after a few days of data collection, once the enumerators become familiar with the survey and are able to reach the daily target. It is recommended to allow some flexibility in the beginning while the teams are familiarising themselves with the tool etc. to avoid compromising data quality in the beginning.
    - Note that it can be an issue if enumerators are conducting both too few or too many interviews.
        - → Too few can indicate that they are not comfortable collecting the data, that they lack motivation or that the planned number did not properly consider travelling time in rural/deserted areas.

→ Too many can indicate that the enumerator is rushing through the interviews and not using enough time to probe for high-quality data.

❑ Check the **geographical information** against the dates of data collection or the name of the enumerator to ensure that households were accurately captured in the correct locations.

❑ Check the **GPS coordinates** to see how random the sample is within the geographical areas, and check for main road bias.

❑ **Respondent IDs:** Check for duplicates of household personal identification information and/or phone numbers. This could be due to accidental revisits, typing mistakes or duplicate submissions.

❑ **Missing values:** If the quality checks show that some key questions have missing values that are not due to logical skip patterns, this must immediately be flagged to the team leader to ensure that the enumerator is using the correct form. If this is a general issue, the Country Office may need to initiate a new, updated form though this is generally not recommended when teams have already travelled to the field.

❑ **Patterns of data entry** such as the same entry repeated many times (7, 7, 7, 7, 7) or alternating numbers (2, 1, 2, 1, 2, 1, 2, 1) by enumerators should be flagged as possible data quality issues.

❑ **Erroneous values:** Look for mixed-up numbers that may arise in the absence of constraints, e.g., a respondent had 10,000 income sources, which is likely it was the value of income, not the number of their sources.

❑ **Skipping modules:** Check the average number of values that prompt a skip module to detect if enumerators seem to be systematically misusing the skipping questions in the survey to save time.

OUTLIERS EXPLAINED: Sometimes, outliers can be explained, e.g., it could be that one team is surveying a deserted pastoralist population which have a diet that differs significantly from the rest of the population, or that data is collected from an area that is either among the most or least food secure. This should be considered before communicating the HFC flags to the team leaders.

❑ Examine responses to '**other, please specify**,' to check that the responses are correctly identified as outside of the pre-existing question list and are clear enough for recoding. Miscoded responses should be flagged and communicated to the team leaders if a value for 'other, please specify' should be recoded as an already existing response option. If the frequent use of 'other, please specify' is found to be correctly used, e.g., if an income source is common but not already included in the income source module, it is recommended to include this as an answer option in future surveys.

Checks by Module

DEMOGRAPHICS AND HOUSEHOLD ROSTER

❑ Check the **number of adults and children** in the household, and whether the automatic calculation based on individual age/sex groups corresponds to the total number provided by the respondent.

❑ Flag households that have **no adults** as potentially erroneously collected data.

❑ Flag households with no/more than one **head of household**.

❑ Flag households with a **total household size** that appears unrealistically high for the context.

❑ Flag households where the **number of IDPs** hosted seems unrealistic for the context.

KEY INDICATORS

Calculate key indicators (FCS, rCSI, LCS, FES), and check the data for outliers, by enumerator and by geographical area. For all the below checks, it is important to note that for the most part, **flags are not necessarily errors, but rather, anomalies to follow up on** – especially if particular enumerators are flagged.

When checking key indicators by enumerator, it is recommended to also check the distribution of the enumerator within the team as this is useful to indicate potential sampling issues or if an outlier can be explained by the geographical context. As an example, the mean FCS should not be too different for the team members covering the same geographical areas considering random sampling. Hence, if one team member is consistently reporting a very low FCS compared to the other team members, ensuring a sufficient sample size before flagging, it could indicate that the enumerator has misunderstood the indicator. If instead the entire team consistently reports very low FCS, it can indicate that the geographical area is experiencing extreme food insecurity.

FOOD CONSUMPTION

❑ **FCS:** The FCS module should be constrained to contain values between 0 to 7 number of days in a week; it is not possible to exceed (>7) or to have negative values.

❑ Check the food consumption module for low and high food consumption of key food groups.

❑ Global thresholds to be used across WFP operations:

- **Low cereal consumption (<=4 days)**, flag unless replaced by consumption of pulses/legumes.
- **Low oil and sugar consumption** in contexts where these groups are normally eaten daily (using adjusted, high FCS threshold).
- **Very high consumption of meat and dairy** (e.g., >=5) in contexts where these groups are normally not frequently eaten as this can indicate issues with understanding of small quantities.
- **Entire module is filled with zeros** (no food consumption in the past 7 days) in contexts that are not expected to be seeing famine-like conditions, triangulate with other indicators (e.g., HHS, rCSI) if they support this. If not, flag to team leader. If famine-like findings are consistent throughout the food security data, flag immediately to the team leader to confirm whether starvation may be happening in the area.
- **FCS was lower than 14**, reflecting a very dire situation where the households did not even consume cereals daily.
- **FCS higher than 100**, meaning that households eat nearly all eight food groups all days of the week.

❑ **FCS-N:** The number of days of consumption of food subgroups can never exceed the number of days the main food group was consumed.

❑ **Non-mandatory checks** that can be used depending on relevance in context:

- **High values of food groups** are usually not consumed often, e.g., fruits, fish in remote/ desert areas etc.
- **Low values of food groups** that are normally relied on in the context, e.g., fish in coastal areas that normally rely on consuming fish during certain seasons.
- Calculate the **average number of consumption days for each food group** disaggregated by enumerators, paying special attention to cereals, pulses, oils and sugar, dairy and protein.

COPING STRATEGIES

❑ **rCSI:** The rCSI strategies should be constrained to contain values between 0 and 7 for the number of days in a week; it is not possible to exceed (>7) or have negative values.

- **rCSI >= 42**, meaning that households had a very high usage of food-based coping.
- **rCSI <=3**, meaning that households barely applied in any food-based coping strategies. Check against FCS, if poor or borderline, flag to team leader.
- Logically inconsistent answers (e.g., strategies involving children for households with no children).

❑ **LCS:**

- For some coping strategies, certain response options are not applicable. For example, it is rarely possible to 'exhaust' begging or theft. However, they can be not applicable in extremely remote areas where there are no households reachable to beg/steal from. Note that this should be distinguished from being considered not socially acceptable.
- High use of N/A for all or most of the LCS questions. Note that while this should be flagged to the team leader, it can also indicate a design issue that the module has not been properly contextualized.
- Logically inconsistent answers (e.g., strategies involving children for households with no children).

HOUSEHOLD HUNGER SCALE

❑ **HHS:** The HHS module should be constrained to contain values between 0 and 3, depending on the number of days each behaviour has been applied; it is not possible to exceed (>3) or have negative values.

- **HHS of 5 or 6**, meaning that households experienced very severe hunger in the past 30 days. If the data was collected in contexts that are not expected to be seeing famine-like conditions, triangulate with other indicators (e.g., FCS, rCSI, LCS) if they support this. If not, flag HHS to team leader. If famine-like findings are consistent throughout the food security data, flag immediately to the team leader to confirm whether starvation may be happening in the area.

EXPENDITURES

❑ **Number of interviews with total food or non-food consumption expenditures of 0.** Observing a household with either no food or no non-food consumption expenditures is very unlikely.

❑ **Check values that do not make economic sense**, e.g., food expenditures of 1 or 2 units when a unit of bread costs 50 in the local currency. If spotted, investigate the extent to which these values occur, and the potential reasons (e.g., misunderstanding with currencies or anomalies with a specific enumerator, or it could be rounding of a monthly expense). These checks should be done with original values.

❑ **Check for outliers.** Use different statistics related to individual expenditure items (and/or aggregates of various kinds) by enumerator/admin, to spot incorrect application of the module. These checks can be done in different ways, but they should all be done in per capita terms and taking into account potential prices differences across survey areas. Examples of possible checks include:

- Comparing mean/median expenditure by enumerator to identify enumerators who might systematically overreport/underreport expenditure. If random sampling is properly followed, there should not be very large mean/median differences between enumerators in the same team.
- Identify outliers using a statistical procedure like that described in the "Cleaning" section and compute occurrences of outliers by enumerator
- Box plots can be a handy way of comparing median, minimum, maximum and extreme expenditure values by the enumerator to detect possible issues with the administration of the module.

❑ **Typos:** While some outliers will naturally look like outliers (e.g., a one-off unforeseen health-related expense), whereby a household has anomalously high expenditures, enumerators might also have included extra zeros by mistake. So, it is best to verify with the field team, where possible.

❑ **Check for atypical values like '33' '77' '88' '99' or other overly specific**, out-of-the-ordinary figures to flag to enumerators and check if they meant not applicable (which is a practice that should not be applied in the expenditure module).

GENERAL DATA CONSISTENCY CHECKS

❑ Additional to the above triangulation, it is recommended to flag data inconsistencies using the 'FEWS NET matrix', which combines HHS, FCS and rCSI. As seen in the example below, each combination of the three indicators has a cell number, and the illogical combinations are considered to be cells number 3, 4, 5, 8, 9, 10. Enumerators with multiple cases falling into one of these six cells should be flagged to the team leader.

❑ Other checks can be added depending on the tool used. Generally, if responses seem inconsistent across modules, this reflects a lack of probing done by enumerators. Please refer to the section above on optional data consistency checks (under Enumerator Training).

Total percentage of unlikely combinations: 0.47%

	rCSI <4			rCSI 4-18			rCSI >18		
	Acceptable FCS	Borderline FCS	Poor FCS	Acceptable FCS	Borderline FCS	Poor FCS	Acceptable FCS	Borderline FCS	Poor FCS
HHS = 0	1 (3.47%)	6 (0.48%)	11 (0%)	16 (3.12%)	21 (1.52%)	26 (0.26%)	31 (1.65%)	36 (0.78%)	41 (0.48%)
HHS = 1	2 (2.47%)	7 (0.04%)	12 (0%)	17 (3.38%)	22 (3.34%)	27 (0.91%)	32 (1.26%)	37 (2.0%)	42 (0.95%)
HHS = 2-3	3 (0.43%)	8 (0.04%)	13 (0%)	18 (11.19%)	23 (8.42%)	28 (1.69%)	33 (8.24%)	38 (19.61%)	43 (17.61%)
HHS = 4	4 (0%)	9 (0%)	14 (0%)	19 (0.04%)	24 (0.04%)	29 (0.09%)	34 (0.13%)	39 (1.91%)	44 (2.26%)
HHS = 5-6	5 (0%)	10 (0%)	15 (0%)	20 (0%)	25 (0.09%)	30 (0%)	35 (0.04%)	40 (1.04%)	45 (1.0%)

ISSUE LOG

All issues identified through the HFCs should be documented and kept. **After communicating to the team leaders, a log should be kept for recording actions to the flags for the upcoming HFCs and cleaning, as well as lessons learnt for future surveys.** This includes:

- Any cleaning needed after the data collection has been finalised when data cannot be corrected in the future. *E.g., Enumerator101 did not ask the FCS module correctly prior to 1/9 2024, so FCS data collected prior to this date should be discarded.*
- Any severe data quality issues that could not be corrected after retraining and multiple efforts by the team leader. *E.g. discard all data from Enumerator202 due to lack of capacity to collect data.*
- Justifications for why some flags are not data quality issues but actual data points. *E.g. food security data collected by Team300 reflects high severity due to a deterioration on the ground.*
- Data reflecting need for better contextualisation of modules. *E.g. livelihood coping strategies are largely not applicable for urban households, and a LCS contextualisation study should be conducted before the next assessment.*
- Modules with many flags, suggesting either that the module does not work well in the context or that the enumerator training needs to be improved. *E.g. household hunger scale reflecting high severity despite not confirmed by the situation on the ground.*

The issue log can be drafted as preferred by the Analyst; some can be noted directly in a syntax file, to be used during the cleaning stage; it can also take the form of a separate document (e.g., Excel or Word file), to be used as a reference during the cleaning stage or as reference for lessons learnt to update the tool/training material for future surveys.

As monitoring the data quality daily is already a time-consuming task, it is recommended to keep the logging burden relatively easy and focus on registering patterns rather than single occurrences, e.g. one enumerator having few issues in one module.

RECOMMENDED ACTIONS FOR MAJOR DATA QUALITY ISSUES

If enumerators are submitting data with a high number of data quality issues, especially in key food security modules, this needs to be flagged immediately to the team leader. The team leader should:

- Stop the enumerator from collecting more data
- Try to understand how the module(s) of concern are being collected and understand if outliers are reflecting the actual situation of the household or are due to lack of enumerator understanding
- If the latter is the case, the enumerator must be re-trained immediately by the team leader, and the team leader should test the knowledge of the enumerator before resuming the data collection.

The analyst conducting the HFCs should pay special attention to this enumerator in the next checks by noting the modules of concern and checking only the data collected after the retraining to assess that no further quality issues are being flagged.

It should be noted that the checks of severe data quality issues can also be requested by the team leader if (s)he is in doubt about an enumerator's performance. In this case, it is recommended for the team leader to ask the Data Analyst to check if the suspected issue is an issue across the data collected.

In case the same quality issues persist, the Country Office and the team leader can ultimately decide to stop the enumerator and use a replacement enumerator from the roster. While this seems like a heavy decision, it should be noted that most often, most data quality issues are coming from very few enumerators. By excluding the few worst performing enumerators, data quality can be significantly improved.

After Data Collection

DATA CLEANING

Data cleaning is the process of identifying and correcting errors or inconsistencies in datasets, ensuring that the data is accurate, complete, and ready for analysis. It involves handling missing values, removing duplicates, and resolving issues identified in the previous phase. **Data cleaning should always be conducted by an experienced Data Analyst after the data collection is finished and before the analysis can start.** While the monitoring of discrepancies and missing data should be an ongoing process through HFCs during data collection, to minimize the number of issues to address or clean at this stage, all remaining issues should nevertheless be identified and resolved at this stage, before moving to analysis and reporting.

During data cleaning, all modules in the entire dataset should be counter-checked, and action should be taken to resolve potential issues. **If the previous steps of coding the questionnaire correctly and conducting daily HFCs are done thoroughly, data cleaning should be minimal.** Furthermore, the data cleaning should be done objectively and systematically, minimizing the level of subjectivity.

FULL CYCLE DATA QUALITY APPROACH: Note that if the previous sections of this guidance have been applied properly during the first phases of the data collection (before and during), such as setting appropriate constraints, warnings, training and testing enumerators well, using HFCs to flag issues, data quality issues should be minimal and therefore very little data cleaning should be expected. However, for the sake of keeping this guidance comprehensive, all possible data cleaning will still be listed.

The Data Analyst should avoid as much as possible making arbitrary decisions regarding cleaning, making sure to apply cleaning procedures in a way that is as methodologically and logically sound, consistent, and in all cases documented while also considering the context where relevant.

Note that all cleaning should always be documented in a syntax/do file, that must be saved in the dedicated script folder (on Data Library), together with the raw and clean dataset. To ease the understanding for other colleagues who may be working on the dataset in the future, the syntax should include the name of the analyst, the month and year of data cleaning as well as notes explaining what decisions were made based on contextual discussions and why. Any cleaning done should be documented so that it is replicable on the same untreated dataset and can serve as a guide to update future questionnaires (including design, coding and training presentations for future surveys).

General Cleaning

MAJOR DATA QUALITY ISSUES

In the case that an enumerator was removed from the data collection due to issues that were not able to be corrected through retraining, all cases collected by the enumerator should be removed from the dataset.

ADHERENCE TO THE SAMPLING PLAN

- ❑ Delete cases where consent was not given
- ❑ Delete ineligible cases, for example, IDP households if not in the inclusion criteria, or geographic areas not covered by the sampling plan.
- ❑ Delete other incomplete cases

DUPLICATE RECORDS

- ❑ Check for duplicate records starting with the unique household identification variable, ID number, and/or phone number.
    - If there are duplicate records (for whatever reason), check if both interviews are complete.
        - → If not, it may be stemming from a submission issue, which can happen especially in areas with connectivity issues.
        - → If both are complete, check if the rest of the households' data are different (e.g., day and time of survey, different figures in the household demographics, ID number, assets, GPS coordinates etc.).
        - → Alternatively, a similarity check should be conducted to identify if there are very similar responses. Check also directly with the field supervisor and enumerators to try to confirm if it was the same household.
    - If they indeed appear to be the **same household** based on either the data and/or confirmation from the field.
        - → If the first interview is complete, then it is recommended to **keep the first interview and drop the second interview.**
        - → If the first interview is incomplete, and the second one is, then **keep the second and drop the first.**
    - If the cases appear to be **substantially different**, (e.g., differing household size, consumption, and other responses), and/or the field is unable to confirm, or indeed confirms that they are two separate households, then it is suggested to **leave them as is.**

SPECIAL VALUES

- ❑ **Special values** (e.g., 9999, 999, 888, 99 to indicate "don't know") do not provide any substantial information for further analysis, and thus should be **replaced as missing** to ensure 'valid percent' results from analyses are reflecting actual values. Refusals and don't know should stay as missing in calculations since no relevant information was provided, thus those responses should be **excluded from the meaningful calculation and interpretation.**

❑ **“Other please specify,” responses** should be **reviewed** in the open-ended follow-up question and **recoded to pre-existing response options or grouped into new response options**, as relevant.

❑ **Be wary of 8s** – as the ‘8’ is located just above the ‘0’ on a tablet/phone keypad, 8s are sometimes a typo for 0s. **Depending on the module and the feedback during the HFCs, the Analyst needs to make a judgement call whether to replace 8s with 0s, code as missing or leave as is.**

Cleaning by Module

GENERAL INFORMATION

❑ Verify that the **geographical information matches up** with the dates of data collection or the name of the enumerator to ensure that households were accurately captured on the correct days in the correct locations. **If not, correct the names of the surveys which were miscategorised. If dates are inaccurate due to a technical issue, code as missing.**

DEMOGRAPHICS AND HOUSEHOLD ROSTER

❑ **Household size:** in cases of unrealistically large households which could be a typo, the analyst can:

- Use the calculated HHSize to check whether the typo comes from a specific sex/age category, e.g., values for all categories are normal while 22 female 60+ are registered. In this case, we can surmise that it should be 2 instead of 22.
- Cross-check against other indicators (e.g., number of rooms in the dwelling)
- If not possible to validate, the analyst can decide to recode HHSize as missing if certain that it is a typo, otherwise, leave it as is.

❑ **Head of household:** If there is more than one head of household indicated by accident, review the cases to try to determine the correct response:

- First address cases that can be determined: e.g., one man aged 45 and one boy aged 12 are indicated; in this case, the 45-year-old man is the household head.
- In cases of doubt, it makes sense to take the older of the two.
- If not possible to determine, then recode any variables related to head of household, e.g., sex, age and education as missing as a last resort, otherwise, leave as is.

FOOD CONSUMPTION

❑ **Food Consumption Score module:**

- The food group indicators should contain values between 0 to 7 number of days in a week; it is not possible to exceed (>7) or to have negative values. **If for whatever reason, some negative numbers or numbers exceed 7, code them and the main FCS indicator as missing, because it is not possible to know what was intended or if it was an error.**
- **Calculate FCS, which should be between 0-112. If not, go back and check, as it was miscalculated.** Incomplete cases should not get a final FCS due to the risk of underestimating the situation of the household. Therefore, **if incomplete code FCS as missing.**
- **For FCS between 0-13**, reflecting cereal consumption of less than 7 days. Unless in contexts that are seeing famine-like conditions, it is very unusual for households to have such low food staple food consumption that they unable to eat even cereal daily. **Check issue by enumerator and triangulate with other related indicators:**
    - → If all flagged cases are coming from the same enumerator(s), this could indicate a data quality issue.

  - → If **all** other food security indicators are pointing towards more food secure households, i.e. no or low coping strategies, especially rCSI (depending on the context), normal/high HDDS, no/low HHS, no emergency livelihood coping and if they spent any money on food, **then replace all food groups under FCS as missing. However, if not possible to rule out that the low FCS could be true, take a no-regret approach and leave it as is.**

- **For FCS between 100-112**, meaning that households eat all/nearly all eight food groups all days of the week. In most contexts, even well-off households will not eat all food groups every day and this could indicate data quality issues. **Check issue by enumerator and triangulate with other indicators:**
  - → If all flagged cases are coming from the same enumerator, this could indicate a data quality issue.
  - → If high coping strategies, especially rCSI (depending on the context), high HHS, low HDDS, **and** emergency livelihood coping **replace all food groups under FCS as missing. However, if not possible to rule out that the high FCS could be true, take a no-regret approach and leave it as is.**

❑ **FCS-N:** Number of days entered for the sub-group should always be less than or equal to the values in the main group. If, for whatever reason, there are households with data in the subgroups that exceeds the maximum values of the main groups, the analyst must decide based **on a contextual discussion of which food group that is easiest to understand in the context whether the main group or the subgroup should be set as the maximum.**

❑ **Food sources:** Check the FCS sources thoroughly for illogical answers (e.g., hunting for fruit, or having assistance as a main food source while not receiving in-kind assistance) by using knowledge of the local context and triangulating with the other relevant data points gathered in the same interview. **Invalid responses should be recoded as missing.**

❑ **Household Dietary Diversity Score module:**

- **Values** must be between 0 and 1; if not, **recode entry as missing**.
- **Missing:** If any value for one of the 12 food groups is missing, leave the response as missing (do not replace it with 0), and it should not be taken into account in the final calculation.
- **Calculate HDDS**, if more than 12, double-check the calculation as it is possible that you included incorrect values.
- **Triangulate with FCS:** If HDDS is 0 and FCS is 7 for any corresponding food group, then for HDDS, replace 0 with 1. Note that reverse cleaning is not possible since HDDS includes small quantities and counts consumption for anyone.

❑ **Household Hunger Scale module:**

- **Values** must be 0 if the household did not experience hunger according to the question asked and between 1-3 if the household says they experienced it, if not **recode entry as missing**.
- **Missing:** If any value for one of the 3 questions is missing, leave the response as missing (do not replace with 0), and it should not be taken into account in the final calculation.
- **For HHS of 5 and 6**, meaning that households experienced severe hunger pointing towards famine-like conditions **triangulate with other indicators and check potential issues by enumerator:**
  - → If no/low coping strategies, especially rCSI (depending on the context), acceptable FCS, high HDDS, **and** no/little livelihood coping **replace HHS as missing. However, if not possible to rule out that the high HHS could be true, always take a no-regret approach and leave it as is**.
  - → If all flagged cases (high HHS not justified by triangulation) are coming from the same enumerator, this could indicate a data quality issue.

COPING STRATEGIES

- ❑ **Reduced Coping Strategies Index module:**
  - **Values:** The rCSI module should contain values between 0 to 7 number of days in a week; it is not possible to exceed (>7) or have negative values. **If for whatever reason, some negative numbers or numbers exceed 7, replace them as missing, because it is not possible to know what was intended or if it was an error.**
  - **Triangulate strategies with relevant modules** depending on what has been included in the questionnaire. This includes ensuring that households that indicate that they have used the strategy of restricting adult consumption for children have listed having children in the demographic section.

- ❑ **Calculate rCSI**, if not between 0-56, double-check the calculation as it is possible that you included incorrect values. **Incomplete cases should not get a final rCSI score** due to the risk of underestimating the situation of the household.

- ❑ **For rCSI >= 42**, reflecting extremely high use of food-based coping, **check by enumerator and triangulate with other related indicators:**
  - If all flagged cases are coming from the same enumerator(s), this could indicate a data quality issue.
  - If **all** other food security indicators are pointing towards more food secure households, i.e. acceptable FCS, normal/high HDDS, no/low HHS, no emergency livelihood coping and if they spent any money on food, **then replace all strategies and rCSI value as missing. However, if not possible to rule out that the high rCSI could be true, take a no-regret approach and leave it as is.**

- ❑ **Livelihood-based Coping Strategies module:**
  - N/A or 9999 remains as is.
  - For some coping strategies, certain response options are not valid/possible. Unless in extremely remote areas, it is rarely possible to 'exhaust' begging or theft. **If these strategies are marked as N/A, then responses should be recoded as "No, I did not apply this strategy."**
  - **For emergency coping**, reflecting extreme asset depletion, **check by enumerator and triangulate with other related indicators.** However, it should be noted that this indicator is difficult to clean due to the long recall period compared to other food security indicators, which is up to 12 months.
  - **Triangulate strategies with relevant modules** depending on what has been included in the questionnaire. This includes triangulating borrowing/debt coping against the credit category in the Expenditure module, education/child-related coping against information about children and age etc.

EXPENDITURES

Expenditure cleaning

The cleaning procedure is contained in a single syntax file, which can be found on GitHub, is divided into the following sections:

- Preliminary checks
- Stage 1: cleaning variable by variable, manual
- Stage 2: cleaning variable by variable, statistical/ automatic
- Stage 3: cleaning aggregates, statistical/ automatic.

Preliminary check

- ❑ The syntax file produces a table indicating the share of observations in your dataset that present either zero total food or zero total non-food consumption expenditures.

❑ Such cases are highly unlikely, if not impossible. Hence, they might safely be interpreted as coming from an invalid administration of the module or non-response.

Code	Name	Count	Proportion
zero_F	Zero Food Expenditures	0	0%
zero_NF	Zero Non-Food Expenditures	3	6%
zero_total	Zero Total Expenditures	0	0%

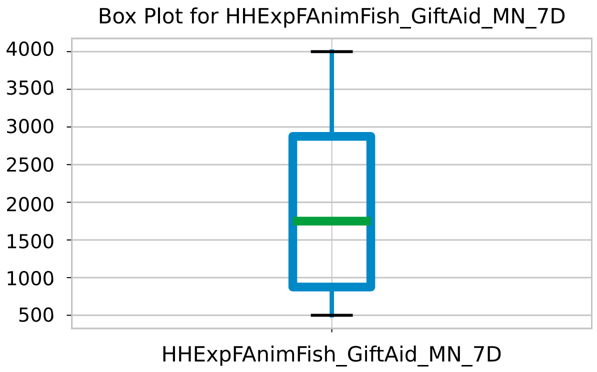
❑ Carefully examine the table. If the share of these problematic observations is very high (e.g., higher than 15 percent), consider the following solutions:

- Excluding from the analysis a certain geographical area, if these problematic cases are concentrated there.
- Excluding cases from problematic enumerators from the analysis, if these problematic cases are concentrated around few enumerators.
- Refrain from using the expenditure module at all.

Stage 1: cleaning variable by variable, manual

❑ In the first stage, cleaning is done **on each single variable** of the expenditure module to identify and manually correct anomalous values. The syntax file will produce:

a. Box plot graphs of each variable in the expenditure module:



Noting that expenditure data is the most complex type of food security data to clean, an automated syntax has been developed. It is recommended to look at the data before and after running the syntax to follow what has been changed, but it is not necessary to understand all statistical procedures or the entire syntax itself. If preferred, all steps can be done manually, however, using the script will save time and reduce risk of errors.

b. A table indicating the bottom and top five values of each variable.

	HHExpFCer_Purch_MN_7D	HHExpFCer_GiftAid_MN_7D	HHExpFCer_Own_MN_7D
Bottom	200	100	300
Bottom	200	100	500
Bottom	200	200	750
Bottom	250	200	1000
Bottom	400	500	1250
	...	...	...
Top	9000	2500	7000
Top	7900	2300	4000
Top	7500	1900	3550
Top	7450	1850	3500
Top	7300	1800	3000

Using the graphs and/or the table produced do the following:

❑ For each variable, look for values that **look implausibly high or low** in the context. Consider for example, the lowest possible unit of currency and the costs of food and non-food items in the country, which should have been discussed during the enumerator training. Check household size to judge on extremely high values.

❑ If you identify implausibly low or high values, **judge whether an obvious mistake occurred**, and if it can safely be replaced with the 'true value'. Examples include cases where enumerators put a different currency to report values; reported in '000 while they were not supposed to; or accidentally entered a negative value.

- If this is the case, **correct the values** in the syntax section 'manual corrections'. Add a note explaining why and how each value is replaced to ensure all cleaning decisions are documented and replicable.

Stage 2: cleaning variable by variable, statistical/automatic

❑ This section of the syntax file identifies any remaining outliers in per capita terms and automatically replaces them with the median value of each variable.

❑ This section of the syntax file is fully automatic, and does not require any manual edits.

Stage 3: cleaning aggregates, statistical/automatic

❑ The final cleaning stage is performed on two aggregates: 1) total food and 2) total non-food consumption expenditures, to identify and correct outliers.

❑ The syntax identifies observations where:

1. **Total food or total non-food expenditures equal zero.**
2. **Outliers**, e.g. ± 3 SD from the median in per capita terms.

❑ The syntax automatically cleans both types of observations by:

- Replacing identified outliers and aggregates that equal zero with the median value of each aggregate. Median values to be used as replacements are calculated at lowest possible geographical level (based on data availability) to better account for price differences.
- **Adjusting the individual variables** of the aggregate to reflect these changes. This is done by allocating the new values of aggregates to the individual variables, based on the average expenditure share of each variable in the population.

Final steps

❑ The syntax provides household level results as well as outputs to track and assess the impact of the cleaning procedure on the data:

- An **Excel file** including:
- A dataset storing the original values of the variables, indicating whether and by how much each data point was edited in the cleaning procedure.
- A summary of share of observations that were edited for each variable.
- Tables showing mean/median expenditure aggregates and food expenditure share, before and after cleaning.
- Box plot and other graphs visualising expenditure aggregates and food expenditure share, before and after cleaning.

❑ These outputs give an overview of the data quality issues associated with your expenditure data. This will also help take note of common errors and take corrective actions for future data collection exercises.

❑ The final output will be a dataset with cleaned expenditure variables that are saved with the standard names of the WFP Codebook.

DATA AND DOCUMENT MANAGEMENT

Ensure to upload all final documents to the assessment folders in Teams/SharePoint created prior to the data collection. Archive any working documents/temporary files that are no longer relevant, meaning that only the final budget, final sampling document, final weighting calculation, final enumerator contracts, final (merged) raw and cleaned datasets, final scripts etc. should be in the folder. This is to avoid confusion for other colleagues who will be using the documents later, e.g., to plan for a new assessment or to use the data for additional analysis.

A final version of the cleaned dataset for analysis should be saved in Data Library or a shared folder with a proper setup on who can see, download, and edit the datasets and other relevant materials. Before you upload or share with external stakeholders (donors, programme partners, etc.), please make sure:

- All the Personally Identifiable Information (PII) is removed or recoded from the published/external versions before publishing as public/external on DataLib. Common PII variables include name, phone numbers, address (village level and below, and other significant location markers, such as zip codes), national ID and beneficiary ID, photos, fingerprints and bank information of respondents and their household members. The goal of the de-identification exercise is to protect the privacy of respondents so that any unauthorized people cannot identify them using the available datasets. Please also review or remove all the open-ended questions, such as other specify and comments fields entered by enumerators manually, where PII might also be included. Additionally, the names of enumerators and supervisors should be recoded to numbers or any other anonymized version. It is suggested to implement a de-identification process as soon as the datasets have been downloaded from MoDa and remove unnecessary personal information in the analysis.

- A separate folder under the 'Data' parent folder has been created and ideally encrypted (with proper password management) to store the raw data with PII variables, especially if panel data collection is planned. The relevant personal information should be pre-populated into the further round of survey and only authorized staff have access to this process.
- The final dataset can be re-created from the raw stage using the accompanying scripts by not only the analyst but also other colleagues who were not involved in the analysis if tasked. The other analyst should be able to get the same output.

- The scripts used for data cleaning and analysis are properly annotated and presented with notes including:

  → Purposes and specific data collection activities, author/analyst contact (team contact if it's shared with external partners), and last change dates of the scripts.

  → Overview of the script structures by sections.

  → Comment enough on the functions of the codes to help others understand.

  → Decision made and rationale of each cleaning step and data points change. For example, when we drop a specific observation, add a note of why it gets removed.

Annex

ENUMERATOR TEST QUESTIONS EXAMPLE

Specific questions should be kept confidential and focus on ensuring especially data quality in key food security indicators. It is recommended that at least 1-2 questions are included for each food security indicator. Examples of formats can be:

General Questions

1. **What is the best practice if a respondent refuses to participate in the survey and complains they didn't receive the promised assistance by WFP? (Select all that applies)**

   a. Convince the respondent to take the survey and promise that assistance will come.

   b. Report to field supervisor.

   c. Record "No" in the consent and leave.

   d. Collect additional details, such as beneficiary ID.

   e. Reach out to programme colleagues to investigate the case.

2. **In the household size question, if the respondent reports living with his wife but has a son living abroad for studies, whose educational fees they cover, what should be recorded as the household size in this case?**

   a) 2

   b) 3

Food Consumption Score Questions

1. **When filling out the Food Consumption Score (FCS) module, what should you do if the respondent reports that their household did not consume any cereals or staple foods in the past 7 days?**

   a) Record 0 days and move on to the next question.

   b) Probe further to ensure the accuracy of the answer, as it is unlikely that a household did not consume any staple foods in the past 7 days.

   c) Record "Not Applicable" and proceed with the survey.

   d) Skip the question since it seems the household does not consume staples.

2. **If a respondent reports consuming meat on 2 different days during the past week, but they consumed meat twice a day on those days, how should this be recorded in the FCS?**

   a) Record 2 days of meat consumption, as the frequency of meals per day does not affect the number of days recorded.

- **b)** Record 4 days of meat consumption since they consumed meat twice per day.
- **c)** Record the number of meals instead of the number of days (4 meals).
- **d)** Skip the question, as the number of meals is too complex to record.

Reduced Coping Strategies Index (rCSI) Questions

1. **If a respondent reports that their household relied on less preferred and less expensive food to cope with lack of food or money for food on 3 days, but later mentions they also borrowed food from neighbors on those same 3 days, what should the enumerator record?**
   - **a)** Record 3 days for each strategy (less preferred food and borrowing food).
   - **b)** Record 6 days for both strategies, combining the days.
   - **c)** Probe further to understand whether these strategies were used on the same or different days and record them separately if needed.
   - **d)** Skip the question if the respondent is unsure.

2. **If a respondent reports restricting consumption by adults in order for small children to eat, but the adults only did this for one meal on a single day in the last week, how should you record this for the rCSI?**
   - **a)** Record 0 days, as it was only for one meal and not for the whole day.
   - **b)** Probe further to clarify how many meals the adults restricted, and average this across the week.
   - **c)** Record 1 day, as restricting for even a single meal counts for the day.
   - **d)** Skip the question, as the answer is unclear.

Livelihood Coping Strategies Questions

1. **If a household member had to migrate temporarily, but this was their regular seasonal migration pattern, how should the enumerator record the answer?**
   - **a.** Record "No, because we did not need to" as it's part of their regular seasonal activities and not due to food insecurity.
   - **b.** Record "Yes" as they did migrate.
   - **c.** Record "Not applicable (don't have access to this strategy)" as it is a regular activity.
   - **d.** Probe further to understand if the migration was primarily driven by food insecurity or was a regular occurrence regardless of food security.

2. **What should be recorded if the respondent mentions reducing expenses on education due to lack of food but also states they have no school-going children?**
   - **a.** Record "Yes" for reducing expenses on education.
   - **b.** Record "No, because we did not need to" since there are no children to spend on.
   - **c.** Probe further to clarify the answer or understand the situation better.
   - **d.** Record "Not applicable (don't have access to this strategy)" since there are no school-going children.

Food Expenditures Questions

1. **If a respondent reports purchasing cereals in cash and on credit in the last 7 days, how should the enumerator record the amount spent?**

   a. Record the total amount spent on cereals in cash, regardless of whether they purchased cereals on credit.

   b. Record the amount spent on cereals in cash separately from the amount spent on cereals on credit.

   c. Only record purchases made with cash and skip those made on credit.

   d. Add both purchases in cash and on credit and record the total.

2. **What should the enumerator do if the respondent reports consuming cereals received as in-kind gifts but is unsure of the value?**

   a) Skip the question since the respondent is unsure of the value.

   b) Make a qualified guess based on your knowledge on potential local prices and record it.

   c) Probe further to help the respondent approximate the value based on the amount consumed.

   d) Record "Not applicable" and move on to the next question.

Non-Food Expenditures Questions

1. **When asking about fuel expenditures, a respondent mentions that they received $10 worth of gasoline from a friend for free, and also purchased $5 worth of gasoline on credit. How should the enumerator record the answer for "Considering both purchases made in cash and on credit, how much did your household spend on fuel in the last 30 days?"**

   a) Record $15 which is the total value of gasoline purchased on credit and the market value of the gasoline received from his friend.

   b) Only record the fuel purchased on credit $5, as in-kind gifts should not be considered under this question.

   c) Record $0 since there was no purchase made in cash.

2. **When asked the question: "In the last 6 months, did your household purchase any or pay for rent, using cash or credit?", the respondent answers that $100 is paid using cash every month. How should this be registered?**

   a) Record $100 and tell the team leader that rent is paid per month in this area, not after 6 months.

   b) Skip the question, as the recall periods are not matching and it is better to be on the safe side.

   c) Multiply the monthly rent by 6 to get the total rent for the 6 months, which is $600.

Mayo/May 2023
24 Miércoles/Wednesday

Acronyms

CARI	Consolidated Approach for Reporting on Food Insecurity
CO	Country Office
CSB	Corn Soya Blend
ECMEN	Economic Capacity to Meet Essential Needs
FCS	Food Consumption Score
FCS-N	Food Consumption Score Nutrition Quality Analysis
FES	Food Expenditure Share
GFA	General Food Assistance
HDDS	Household Dietary Diversity Score
HFCs	High Frequency Checks
HHH	Head of household
HHS	Household Hunger Scale
IPC	Integrated Food Security Phase Classification
LCS	Livelihood Coping Strategies
MUAC	Mid-upper arm circumference
rCSI	Reduced Coping Strategies Index
VAM	Vulnerability Analysis and Mapping
WFP	World Food Programme

WFP
World Food Programme
ÇA DEMENAGE
Programa Mundial de Alimentos
Programme Alimentaire Mondial
WFP remer
sa con
UNITED NATIONS UNIES

Photo credits

Cover:	WFP/Imtiaz Mahbub Mumit
Page 2:	WFP/Marianne Jensby
Page 4:	WFP/Giulio d’Adamo
Page 6:	WFP/Marianne Jensby
Page 11:	WFP/Marianne Jensby
Page 25:	WFP/Abood al Sayd
Page 26:	WFP/Rein Skullerud
Page 32:	WFP/Kevin Gitonga
Page 34:	WFP/Marianne Jensby
Page 42:	WFP/Irshad Khan
Page 46:	WFP/Nelson Pacheco
Page 48:	WFP/Charlie Musoka

World Food Programme

APP Food Security & Nutrition Analysis Service
Needs Assessment & Targeting Unit

Via Cesare Giulio Viola 68/70,
00148 Rome, Italy - T +39 06 65131

Global.ResearchAssessmentMonitoring@wfp.org