Child, Caregiver & Household Well-being Survey Tools for Orphans & Vulnerable Children Programs:

Data Management Guidance

MEASURE Evaluation
MANUAL

Cover photograph by Zahra Reynolds, MEASURE Evaluation, of children at play in Liberia.

Child, Caregiver & Household Well-being Survey Tools for Orphans & Vulnerable Children Programs: Data Management Guidance

This publication has been supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement GHA-A-00-08-00003-00, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, with Futures Group, ICF International, John Snow, Inc., Management Sciences for Health, and Tulane University. The views expressed in this publication do not necessarily reflect the views of PEPFAR, USAID or the United States government.

July 2014 MS-14-89

Acknowledgments

This data management guidance was developed by Mona Mehta Steffen (consultant) with input from Dr. Lisa Parker, Dr. Jenifer Chapman, and Dr. Zulfiya Chariyeva. This guidance was reviewed by Dr. Janet Shriberg at USAID. This document was edited by Margo Young (consultant) and formatted by Nash Herndon at MEASURE Evaluation.

The development of this toolkit was highly participatory. Materials represent the current best practice around the measurement of orphans and vulnerable children (OVC) and caregiver well-being in the context of the U.S. President's Emergency Plan for AIDS Relief (PEPFAR)-funded OVC programs. At USAID, the development of this toolkit was shepherded by Dr. Janet Shriberg and Gretchen Bachman, with key input from the wider PEPFAR Orphans and Vulnerable Children Technical Working Group, especially, Dr. Beverly Nyberg at Peace Corps and Dr. Nicole Benham at the Office of the Global AIDS Coordinator. We thank Dr. Krista Stewart at USAID for her guidance as the MEASURE Evaluation agreement officer representative.

The important contributions of implementing partners, researchers, government staff, and other stakeholders, too numerous to list, cannot be overstated. This is truly a community toolkit, and we are grateful to our colleagues for their generosity of time, resources and experience.

Table of Contents

Measuring OVC Outcomes: A Toolkit

MEASURE Evaluation has produced a set of tools for measuring quantitative child outcomes and caregiver/household outcomes in programs for orphans and vulnerable children (OVC). This toolkit was developed with the support of the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) Orphans and Vulnerable Children Technical Working Group to:

- standardize the production of population-level child and caregiver well-being data beyond what is available from routine surveys;
- produce actionable data to inform programs and enable mid-course corrections; and
- enable comparative assessments of child and caregiver well-being and household economic status across a diverse set of interventions and geographical regions.

Who Will Use the Toolkit?

The tools will support OVC programs and research institutions with an evaluation agenda. The toolkit may be useful to you if are seeking to answer one of these five questions:

1. Is my program having, or did my program have, an impact on the children and households it reached?
2. What are the characteristics of children and their caregivers in my area regarding education, health, protection, and psychosocial status?
3. Where do the children most in need of program support live?
4. Approximately how many children need services or support?
5. What are the needs of my program's registered beneficiaries as an entirety, in terms of education, health, protection, and psychosocial support?

While no single data collection tool can meet all OVC data needs, this set of survey tools responds to distinct information needs related to program planning and evaluation. These tools will help to standardize measures and processes for assessing child, caregiver, and household well-being at the population level.

Tools in the Toolkit

The toolkit is available at:

http://www.cpc.unc.edu/measure/our-work/ovc/ovc-program-evaluation-tool-kit.

The toolkit includes:

- *Child, Caregiver & Household Well-being Survey Tools for Orphans & Vulnerable Children Programs: Manual* and three questionnaires: Caregiver Questionnaire, Child Questionnaire Ages 0-9 Years, and Child Questionnaire Ages 10-17 Years;
- Child, Caregiver & Household Well-being Survey Tools for Orphans & Vulnerable Children Programs: Protocol Template;
- Core OVC Program Impact Indicators;
- Data Collectors' Training Manual and Materials;
- Data Analysis Guide; and

- Data Management Guidance (*this document*);

Manual: The toolkit's manual describes the tools, question by question, and outlines how the tools may be used as well as how they should *not* be used. The manual also includes basic guidance on implementing the tools, such as the following:

- Program outcome data should be collected by trained data collectors external to service delivery.
- A documented research protocol, outlining a technically robust, peer-reviewed study, is required.
- The protocol, including data collection tools, must undergo ethical review in the country of research.
- Tools should be pilot-tested in the research setting.

Protocol Template: A research protocol is a prerequisite to implementing the OVC questionnaires. The process of protocol development facilitates agreement on the implementation strategy and child protection issues, among other things. This process improves the study design, enables matching of resources to objectives, and ultimately improves the usability of the data generated from the study. Finally, the protocol is a guidance document for all stakeholders throughout the study period, serving as a reminder to all stakeholders of the agreed strategy and timeline. The protocol template has been structured to resemble an actual research protocol, including appendices. For each section, the information that is required has been outlined, as well as issues to consider when developing your own protocol.

Data Collectors' Training Manual and Materials: The field workers who will seek informed consent and administer these questionnaires must be well trained before data collection begins. MEASURE Evaluation developed a data collectors' training manual and materials to ensure a standard level of competency across data collectors. This manual describes the structure and content for a six-day training (five days for data collectors and one additional day for supervisors). MEASURE Evaluation has also developed Microsoft PowerPoint slides and handouts for the training, included with the manual.

Data Analysis Guide: Developing a data analysis plan is the first step in data analysis. A data analysis plan is important as it enables discussion and agreement of the key points of analysis, ensures that the analysis plan will address the research questions, and ensures that the analysis reflects the strengths and limitations of the data. To reduce the burden on evaluators, MEASURE Evaluation developed a data analysis guide for the child, caregiver and household well-being survey tools. This guide includes suggested analyses for the toolkit questions, as well as more general background and refresher information on data analysis.

Data Management Guidance: This is to be used to guide the data management of the OVC questionnaires (*this document*). The purpose of this guide is to: standardize data management procedures; outline the steps for database design; describe best practices in data entry and data cleaning; identify where data-related missteps can occur; and highlight the importance of proactive data management.

1. Introduction

This data management guidance is a part of the PEPFAR orphans and other vulnerable children (OVC) program evaluation toolkit developed by MEASURE Evaluation. For the purposes of this document, when we say data management, we refer to the following procedures:

- Database design
- Data entry
- Data cleaning

1.1. Purpose

This document is to be used to guide the data management of the *Child, Caregiver & Household Well-being Survey Tools for Orphans & Vulnerable Children Programs.*

The purpose of this guide is to:

- standardize data management procedures;
- outline the steps for database design;
- describe best practices in data entry and data cleaning;
- identify where data-related missteps can occur; and
- highlight the importance of proactive data management

Guidance on data management is necessary to ensure that the data are standardized and effectively managed for optimal analysis.

1.2. Description and Audience

This guidance document outlines how data management can be conducted in such a way that supports data consistency and improves data quality. The audience for this document is program staff who will use MEASURE Evaluation's child, caregiver, and household survey tools for data collection. Data collection partners in the field can use this guidance to inform both enumerator and data entry training.

1.3. Structure and Content

This guidance covers the three questionnaires: (1) OVC caregivers; (2) Children aged 0-9 years; and (3) Children aged 10-17 years. We present specific steps for database design, data entry, and data cleaning. For each of these questionnaires, we provide recommendations following the sequence of questions for each questionnaire.

2. Guidance for Data Management

Effective data management is critical to ensure that results from the child, caregiver, and household survey tools are accurately captured and is a step toward achieving and presenting reliable and error-free results. As donors and developing country governments tighten funding, resources earmarked for primary data collection can go farther when data management guidance is available and practiced.

It is crucial to have someone responsible for overseeing data management who understands all aspects of data collection, what to look for to ensure data quality, and how to address data issues.

2.1. Data Management Includes

- Database design: Database design covers variable and data field definitions, data standardization, and data manipulation required to proceed to data analysis.
- Data entry: Data entry is the process of transcribing information (for instance from completed questionnaires) into a database.
- Data cleaning: Data cleaning refers to detecting and correcting data deemed inaccurate. Data entered that is nonsensical must be corrected prior to analysis so results do not generate inaccurate or skewed outcomes. For example, an age value for a head of household of 4,000 is clearly nonsensical and would need to be corrected to avoid skewing the mean head of household age. Another example of data cleaning is replacing the following combinations with the word "street" in an alphanumeric field called address: str, st, stre.
- Data processes: Specific procedures to ensure data safety and security. The database should be stored on a hard drive with a password to prevent data loss or alteration. The database should be backed up daily so lost data are retrievable. The final database and supporting documents such as questionnaires and reports should be archived until the project's end.

2.1.1. Database Design

For effective database design, databases should be developed in a software package such as CSPro, ACCESS, EpiInfo or SPSS. These programs allow skip patterns across variables, specify permitted data ranges and values to track questionnaire instructions, and establish another layer of data quality. We urge all programs to use these capabilities when entering data in CSPro, ACCESS, EpiInfo or SPSS.

Database design for each questionnaire should follow the questionnaire structure, with intuitive and understandable variable names that are easily lined up with questions included on the questionnaire.

Specific components that are important to strengthen database design are discussed below and include the following:

- Naming of Database Files
- Data Dictionary
- Key Variables and Variable Order

- Database Design for the OVC Household Schedule
- Data Standardization
- Skip and Fill Patterns
- Multi-Response Questions
- The "Other" Response

Naming of Database Files — It might be helpful to consider using standardized ways of naming the database file documents. Database names could include the country name, sampling focus (OVC), the respondent group (caregiver, children aged 0-9 years, children aged 10-17 years), year of data collection, and the type of survey (e.g., pilot, baseline, midline, endline). For a baseline effort of OVC in Kenya, we would recommend the following database name: KenyaOVC_Caregiver_2014_Baseline.

Data Dictionary — The tables below show examples of a data dictionary generated in SPSS. Table 1 shows the file name and location, variable names, variable order, and other data-related information. For example, Variable C002 contains the label "State," which is information on the state where the household is located. The variable is second in the database, it is nominal, and it is alphanumeric so it will be written out (A45). A45 indicates that this is an alphanumeric variable and 45 characters are permitted for data entry. A nominal variable has categories with no intrinsic ranking (for example, state or religious affiliation). Variable "Type_of_Location" is a scale which means it is numeric, and F2 defined as fixed in length with two spaces allocated for data entry.

Variable labels are definitions which reside with the variable and provide more information than what is provided by a (normally) 8 character variable name. Examples of variable labels in the table below are C002 which is "State," and "Type_of_Location," which is "Location Urban or Rural." Without these variable labels, it would be difficult for the data analyst to always discern what results are shown.

Table 1. Data Dictionary

Dataset [1] 'D:\Nigeria_Children0_9_2013_Pilot.sav'

Variable	Position	Label	Measurement Level	Write Format
Child_ID	1	Child Identification	Scale	F8
C002	2	State	Nominal	A45
C003	3	LGA	Nominal	A60
C004	4	Locality	Nominal	A60
C005	5	Village	Nominal	A60
Type_of_Location	6	Location Urban/ Rural	Scale	F2
GPS_ID	7	GPS ID	Scale	F5
WAYPOINT	8	Waypoint	Scale	F5
C007A	9	Latitude	Scale	A24
C007B	10	Longitude	Scale	A24

Value labels are associated with values in close-ended questions, and they are essential for data analysis. The next table provides the variable values and their associated labels. For example, variable C103 has value 1 for FEMALE and value 2 for MALE. Without these labels, the analyst would not know what 1 and 2 represent in analysis output.

Table 2. Data Dictionary (Continued)

Variable Values

Value		Label
TYPE_OF_LOCATION	1	Urban
	2	Rural
RESULT_CODE	1	Completed;
	2	Completed interview but child not home appointment made to return for anthropometry
	3	appointment made for later today with caregiver
	4	appointment made for another day with caregiver
	5	caregiver refused interview
C103	1	FEMALE
	2	MALE
C106	1	Excellent
	2	Very good
	3	Good
	4	Fair
	5	Poor

Key Variables and Variable Order — Key variables are variables which permit file matching and merging. Files will need to be combined for effective data analysis. For example, information about the caregiver such as employment or age may be used in an analysis of education levels among children aged 10-17, based on the hypotheses that caregivers with higher-paying employment may encourage higher education levels among children. These cross-file analyses necessitate merging the caregiver and child database files. To merge files, the same key variable must exist across files that will be merged. Proactively considering how files and data will fit together for analysis will facilitate analysis and can be included in your study's data analysis plan. This step is often neglected during the database design stage.

For the child, caregiver, and household survey tools, every database file should have the variables shown below to enable file matching. The variable types and length must also be identical for these variables in each file to permit merging.

- Household ID
- MemberID from Household Composition Database file

To ensure data quality, each database file should have the same variable order as shown on the survey tool. If the data entry clerk has to enter data out of order from what is shown on the tool, there will be data entry errors.

In Table 3 below, the corresponding variable names and orders should be exact. This table shows a portion of the questionnaire for children aged 0-9 years along with a snapshot of the corresponding database file, which shows the variable order and variable name for data entry and data analysis. Note that we could also see the database variable order through the Data Dictionary.

Table 3. Variable Order

No.	Question	Coding Category		SKIP
C110	Has (child's name) had diarrhea in the last 2 weeks?	Yes No	1 2	If No: C114
C111	Did you seek treatment for the diarrhea from any source?	Yes No	1 2	If No: C113
C112	Where did you seek the treatment from?	Public hospital/health center/post Private hospital/health center Pharmacist Private doctor Community-based agent Shop Traditional health provider Pharmacist Other: _____	1 2 3 4 5 6 7 8 66	
C113	Was (child's name) given any of the following to drink at any time since he/she started having the diarrhea?		Yes / No	
		a) A fluid made from a special packet?	1 / 2	
		b) A pre-packaged ORS liquid?	1 / 2	
		c) A government		

	C110	C111	C112	C113
1	2	.	2	1
2	1	2	1	.
3	1	1	2	88
4	2	.	1	.
5	2	.	1	.
6	1	2	2	2
7	1	1	1	.
8	1	1	2	1
9	2	.	1	.
10	2	.	1	.
11	1	1	1	.
12	1	1	1	.
13	1	1	1	.
14	1	2	1	.
15	1	1	1	.
16	1	1	1	.
17	2	.	1	.
18	2	.	1	.
19	2	.	2	1
20	1	2	2	2

Database Design for the OVC Household Schedule — Because databases are horizontal by design, we urge child, caregiver, and household survey tool users to design the Household Schedule Section of the caregiver questionnaire as an entirely separate database file.

Table 4 shows data for 11 of 20 cases in the database. Each horizontal line represents one entity. At the household level, each horizontal line would represent one household, making it very challenging to undertake household composition distributions because each household member would be entered farther and farther to the right as a new variable in the horizontal database. As an example, suppose that during database design, we use the "largest" household, with 12 members, to determine how many variables to generate for the household composition section. To do this using this horizontal format, we would generate 12 variables to capture this composition. However, households with fewer than 12 members would have many empty variables. Importantly, more empty variables may increase the likelihood of data entry errors.

Table 4. Household Composition Table (Horizontal)

HHID	Relation1_ HHHead	Name1	Member ID (Line from Survey Tool)	Age 1	Relation2_HH Head	Name2	Age 2
1	1	Mercy	1	33	2	Enoch	30
2	1	Faith	1	43	2	Jules	49
3	3	Grace	1	1	4	Carl	6
4	5	Wangu	1	1	3	Pius	4
5	1	Moses	1	22	2	Sami	22
6	1	Peter	1	53	2	Charla	45
7	2	Esther	1	42	1	Kenny	9
8	6	Jacob	1	3	5	Donald	10
9	1	Nandi	1	63	2	Mark	62
10	3	Hope	1	1	4	Fanta	11
11	4	James	1	3	3	Ella	8

Table 5 shows how the Household Composition database file should be designed. Rather than each household horizontally strung out, the data are entered at the member level so one household may be in the database several times, depending on how many members reside there. For example, HHID 1 has seven members while HHID 2 has three members. Given the importance of the household composition data, we strongly recommend this section be entered as shown in Table 5. This format also reduces the number of potentially "empty" variables in the database.

Table 5. Household Composition Table (Vertical)

HHID	Relation_HHHead	Name	Member ID (Line from Survey Tool)	Age	NameLiveHere	NameStayLastNight
1	1	Mercy	1	33	Yes	Yes
1	2	Enoch	2	30	Yes	No
1	3	Marci	3	12	Yes	Yes
1	3	Anne	4	10	Yes	Yes
1	4	David	5	8	Yes	Yes
1	3	Jairus	6	14	Yes	Yes
1	5	Baba	7	1	Yes	Yes
2	1	Faith	1	43	Yes	Yes
2	2	Jules	2	49	Yes	No
2	5	Carl	3	15	Yes	Yes
3	3	Grace	4	1	Yes	Yes
3	4	Carl	5	6	Yes	Yes
3	2	Eunace	6	25	Yes	Yes
3	1	Justice	7	27	Yes	Yes
3	7	Ezekiah	8	60	Yes	Yes
3	7	Patrice	9	60	Yes	Yes
3	9	Willie	10	12	Yes	Yes
4	5	Wangu	1	1	Yes	No
4	3	Pius	2	4	Yes	Yes
4	1	Sylvester	3	40	Yes	Yes
4	2	Patty	4	38	Yes	Yes
4	8	Peace	5	35	Yes	Yes
4	7	Betty	6	65	Yes	Yes

Data Standardization — Values for close-ended questions should also be standardized, thus having the same labels across the database files for all close-ended questions. For example, MEASURE DHS datasets consistently use the value 99 for “Don’t Know” and the value 98 for ”Not Applicable.” Other examples of data standardization include: 1 for Yes, 2 for No; or for sex, 1 for Female and 2 for Male. These small steps during the database design stage will ensure better data quality.

Skip and Fill Patterns — Skip and fill patterns in database design are helpful when follow-on questions are left empty due to a response provided in an earlier question. In the example shown below, if the child did not have diarrhea, the next 3 questions are skipped.

Table 6. Initial Question

No.	Question	Coding Category		SKIP
C110	Has (child's name) had diarrhea in the last 2 weeks?	Yes No	1 2	 If No: C114

To simplify data entry, the database designer could create skip and fill code such that when 2 is entered for Variable C110, the program enters 98 'Not applicable' for Variables C111, C112 and C113.

Table 7. Example of Skip and Fill Based on Initial Question

C110	C111	C112	C113	C114
1	2	1	1	1
2	98	98	98	1

Corresponding to skip and fill data entry design are the concepts of "user missing" and "system missing." These distinctions are generated for data that are missing in certain variables. "User missing" values are defined during the database design stage to capture non-response such as 98 for "Not Applicable." "System missing," on the other hand, is empty cells in the database where we don't know the reason they are empty; data were not entered and the database just considers them to be missing. The data analyst will need to recode "system missing" values into some category such as "response not provided."

The table below shows the distribution for question C110 discussed above. Note that the information to the right of the variable name (C110) is the variable label so the analyst knows what data and question the distribution is for. We see that eight children in the sample had diarrhea in the previous two-week period. Again, we would not see the "Yes," "No," and "Not applicable" without the value labels being defined.

Table 8. Crosstabs

C110 Child had diarrhea in past 2 weeks?	
Yes	8
No	2
Not applicable	2
Total	12

Among these children, we would then like to know how many were treated for diarrhea symptoms. The table below presents the distribution for this analysis. We see that four children received treatment while four did not. The two children who did not have diarrhea show "Not

Applicable." This is an example of "User Missing," as the missing responses were pre-coded during the database design stage.

Table 9. Crosstabs of Follow-up Question

C111 Did you seek treatment for the diarrhea?	
Yes	4
No	4
Not Applicable	4
Total	12

The example in the table below shows "System Missing." We have ten children of school age in our sample but only have seven responses in the distribution. It is unclear at this stage what happened to the other three children.

Table 10. System Missing

C202 Does the child attend school?	
Yes	5
No	2
Total	7

Multi-response Questions — These are questions which can receive more than one response as in Table 11 below. One respondent could check every option provided for the question. These response categories make database design tricky. Some statistical software packages permit multi-response analyses. However, problems during data entry often arise with multi-response questions.

Table 11. Multi-response Question Example

C129 Where did you seek treatment?	
1	Chemist
2	Public Health Clinic
3	Private Health Clinic
4	Private Doctor

The simplest way to set up multi-response questions during database design is to generate a unique binary variable for each response, such as in the following examples:

- C129Chemist (Numeric): 1= Yes 2=No
- C129PublicHealth (Numeric): 1= Yes 2=No
- C129PrivateHealth (Numeric): 1= Yes 2=No
- C129PrivateDoctor (Numeric): 1= Yes 2=No

The "Other" Response — Many close-ended questions include the category of "Other" so respondents are not limited by the responses shown on the questionnaire. During database design, questions that include "Other" should have corresponding variables defined. In the example shown in the table below, the questionnaire shows five options for the question, "Where did you seek treatment?"

Table 12. Example of "Other" Response

C129	Where did you seek treatment?
1	Chemist
2	Public Health Clinic
3	Private Health Clinic
4	Private Doctor
66	Other (Specify):

Variables for the "Other" category should be named so it is obvious which variables they correspond with. For example, we created three variables to match three responses received for the question in the table above. We could name these: C129Other_1, C129Other_2, C129Other_3. The variable names inform the data analyst that these "Other" variables are part of question C129. "Other" variables should be alphanumeric to permit text entry. The total number of "Other" variables generated depends on the total number of "Other" responses received.

Conclusions — Sufficient time should be dedicated to designing a high quality database for each questionnaire. Dummy data could also be used to "test" the efficacy and efficiency of the databases.

2.2. Data Entry

2.2.1. Approach to Data Entry

After the databases are designed, data entry can commence. There will be four OVC questionnaire database files:

- Household Composition
- Caregiver
- Children aged 0-9 years
- Children aged 10-17 years

Questions in the primary caregiver questionnaire start with the letter "P," and the database should have a similar naming scheme. Questions in the child (0-9) questionnaire start with the letter "C," and the database should have a similar naming scheme. Questions in the child (10-17) questionnaire begin with the letter "Y," so corresponding variables should also follow suit.

With a strong database design, data entry should be straightforward. During data entry for all OVC databases files, the data entry clerk should follow the cursor as it moves from field to field entering values as they are shown in the completed questionnaire.

Double data entry is considered the gold standard, time and resources permitting. Researchers should enter at least 10 percent of the data twice, checking for errors, ensuring there is only a 3-5% error rate.

Data entry staff should give extra attention to key identification variables (HHID, MemberID) as well as variables with follow-up questions. Ensure that all data shown in the OVC questionnaire are entered, including information such as Identification Data or Enumerator Name. If the field is on the questionnaire, it MUST be entered in the database. No data should be omitted from data entry.

We recommend a data entry training session to ensure good quality data. Data entry clerks should be made aware of, and trained on, completing skip and fill questions as well as binary variables for multi-response questions. Practice on data entry for questions with an "Other" response is also recommended. Additionally, data entry staff should attend the data enumerator training to familiarize themselves with the questionnaires.

The data entry supervisor should be familiar with the software used for data entry. He/she should also check a certain number (or percentage) of cases entered by each data entry clerk for consistency and accuracy. In a situation where data adjustments need to be made, the data entry supervisor must be consulted.

During data entry training, the clerks should be advised to complete entry of a questionnaire before taking a break as it is difficult to remember where to resume. Clerks should also note on the front page of the questionnaire when the full questionnaire has been entered. Clerks should confirm that the ID on the screen matches the ID on the questionnaire. Finally, the hard copies of the questionnaires should be stored and organized so that staff can check them if any issues arise later on during analysis.

2.3. Data Cleaning

It is important to ensure that data are cleaned, i.e., that they are error-free. We have seen examples of how data can be inaccurate. This section covers how to identify data errors so that they can be corrected prior to analysis.

2.3.1 Guidance for Data Cleaning

The first step for data cleaning is to run a frequency command on each variable in the database. In the example given in Table 13 below, 12 children said they have a birth certificate, and therefore the follow-up question (Table 14) also has 12 responses.

Table 13. Does [NAME] Have a Birth Certificate?

Frequency		
Valid	YES	12
	NO	8
	Total	20

Table 14. Could You Please Show Me [NAME's] Birth Certificate?

Frequency		
Valid	Seen / confirmed	8
	Not seen / not confirmed	4
	Not applicable	8
	Total	20

The second step is to review the frequency results for the following:

- The variables match the order shown in the tool.
- All variables in the database contain data.
- There are no missing values.
- There are no outliers, nonsensical, or unexplained values for variables.
- The number of responses line up across related and follow-up questions.

If outliers, nonsensical, or unexplained values are found, refer back to the original completed questionnaire to confirm what is entered for that question to determine whether the issue arises from an enumerator error or a data entry error. If it is an enumerator error, discuss the situation with the database supervisor to determine how to proceed. All errors and inconsistencies (with corrections) should be documented for future reference. The original uncleaned data file should remain intact and accessible for future reference.

Missing values are commonplace. It is essential to first determine why the data are missing. Sometimes the missing value can be replaced with "Don't Know" or "Not Applicable." Under no circumstance should a missing value be replaced with a "0." Zero is a valid value and will disrupt the analysis if entered inappropriately. If no reason is identified for the missing value, it should remain missing.

2.3.2. Guidance for Cleaning Data from Each Questionnaire

While all data in the database are important and must be accurate and consistent (clean), we highlight specific sections below which may present data entry and quality challenges due to the complexity of the sections.

Caregiver Questionnaire — Data from the food security sections (partially shown) in Table 15 below can be difficult to enter because all the questions are related. Furthermore, some of the questions involve probing. Confirm that the denominators match up and there are no missing values. Then make certain that the skips are correct.

Table 15. Food Security Section

P324	Approximately how much money did your household spend on food in the last one month?	_____Naira	
P325	Was this more or less than the month before, or about the same?	More 1 Less 2 About the same 3	**If MoreP326** **If Less P327** **If SameP328**
P326	Why did you spend more on food? **Probe:** Anything other reason? **Multiple responses ok. Circle all mentioned.**	More people in household now 1 Reduced household food stores 2 Had more disposable income 3 Food prices went up 4 Other: _____ 66	

Ensure that the child age filter variables match the original child age variable throughout the questionnaire. Because the variables are far apart in the questionnaire, enumerators may forget what was entered in an earlier section. In the example presented below of frequency analyses, the enumerator entered the child aged 1 from the initial question as being aged 2 in a subsequent variable. Based on Table 16, there are 19 children aged 2 and older but there is a data entry error in the next table, as it shows 20 children aged 2 and older. When comparing these two frequency results, we should refer back to the original questionnaire to confirm whether the error is data entry or enumerator error.

Table 16. Child Age Distribution

Frequency		
Valid	1	1
	2	3
	3	1
	4	1
	5	2
	6	3
	8	3
	9	6
	Total	20

Instructions for Age of Child question: Please enter 0 if the child is under age 1 and enter value 1 if the child is 2 or over. Table 17 below shows the frequency of children 2 years or older.

Table 17. Age of Child

Frequency		
1	2 YEARS OR OLDER	20

Child questionnaire (aged 0-9 years) — The child's age is recorded in several places. Verify that the age value is consistent everywhere. The child's birth date is a field in the tool, as is the child's age. Confirm that both of these values match, and that the filter variables referred to are also consistent.

Double-check that responses in the school attendance section are consistent, as there are many sub-level questions that build on one another. For example, variable labels for education-related questions in the questionnaire could include the following:

- What grade/class is [child's name] currently in?
- What grade/class was [child's name] in last year?
- Has [child's name] ever attended school?
- When was the last time [child's name] regularly attended school? Would you say it was less than a year or more than a year ago?
- What is the highest grade/form/year that [child's name] has completed?

Check that Question C214, which asks about the engagement of pre-school children with family members, is completed based on the age of sample children. Again, the number of responses for this question should match up with the age distribution from an earlier section in the questionnaire.

Child questionnaire (aged 10-17 years) — Confirm that the ages match across these two age variables as seen in Table 18 below.

Table 18. Match Age Variables

Y104	In what month and year were you born?	Month [__|__] Year [__|__|__|__]	
Y105	How old were you at your last birthday? **Confirm with Y104 and adjust if necessary. Do not leave blank. If child does not know, ask caregiver to estimate age of child.**	[__|__]	

Ensure that the age-specific sections such as AIDS knowledge and sexual activity shown in Table 19 below (partially shown) are completed correctly by the appropriate age group.

Questionnaire instructions specify that this section should be completed by children aged 12 and over, so enumerators should take care that this has been done accurately.

Table 19. Section 7: HIV/AIDS Knowledge, Attitudes & Sexual Behavior

THIS SECTION IS RESTRICTED TO YOUTH AGED 12-17 ONLY.
AGE OF RESPONDENT: [__|__] YEARS – SEE Y105 FOR AGE
IF AGED 10 OR 11 PROCEED TO Y801 A-E

The frequency tables below show that the first question of the section on HIV/AIDS Knowledge, Attitudes and Sexual Behavior is completed incorrectly. According to the initial age question (Table 20), 16 children are aged 12 and over. However, the first question (Table 21) of this section, "Have you ever heard of an illness called AIDS?" presents responses from seventeen children.

Table 20. How Old Were You on Your Last Birthday?

	Age	Frequency
Valid	10	2
	11	2
	12	2
	13	3
	14	4
	15	1
	16	2
	17	4
	Total	20

Table 21. Have You Ever Heard of an Illness Called AIDS?

Frequency		
Valid	YES	16
	NO	1
	Total	17

As discussed earlier, the food security section requires clear focus and understanding because many questions are phrased similarly. The list of variables in this section include the following:

- In the past 1 month, did you have to skip a meal because there was not enough food? How many times did this happen?
- In the past 1 month/4 weeks/2 weeks did you go to sleep at night hungry because there was not enough food to eat? How many times did this happen?
- In the past 1 month/4 weeks/2 weeks did you go a whole day and night without eating anything because there was not enough food? How many times did this happen?

Confirm that all data are entered and consistent across related questions in the food security section. The example below in Table 22 shows that seven respondents skipped a meal due to insufficient food. The follow-up question should also be answered by seven respondents, but only six respondents provided an answer as seen in Table 23.

Table 22. In the Past 1 Month, Did You Have to Skip a Meal Because There Was Not Enough Good?

Frequency		
Valid	YES	7
	NO	13
	Total	20

Table 23. How Many Times Did this Happen?

Frequency		
Valid	Rarely (1-2 times in past 1 month)	2
	Sometimes (3-10 times in past 1 month)	4
	Total	6
Missing	System	14
Total		20

It is also important to check the school attendance section for consistency, as mentioned in the instructions for the questionnaire for children aged 0-9 years.

Final Thoughts — Data management goes a long way toward ensuring clean, consistent, and usable data. Following the steps and recommendations outlined in this guidance will help ensure data consistency and improve data quality.