Three steps to

High Quality Data

By: Monitoring Unit, Regional Bureau Cairo

Step 1: Tool design

- Questions should be organized in logical order
  - Answers should be logical and number-coded
  - Include only questions you need and intend to analyze & report
  - All mandatory indicators should be included
  - Minimize unnecessary open ended questions
  - Sensitive questions should be at the end of the tool

- Pilot test the tool to identify issues before actually data collection. During piloting, answer the following:
  - Do questions measure what it is intended to measure?
  - Is the wording clear/accurate?
  - Do all respondents interpret the question in the same way?
  - Are the questions biased and leading to certain answers?
  - If a telephone survey, do the questions flow in a conversational manner?
  - How long does it take to complete the questionnaire?
  - Does it collect all the information you want to analyze/report later?

Validation rules for mobile tools

Mobile tools shouldn't accept empty fields (if mandatory) and shouldn't accept out of range entries

- Number of children going to school <=[1] number of children in HH
- Total number of HH members = Male + female members
- Food consumption questions accept numbers between (0 and 7)
- If food consumption frequency 'zero', food source should be automatically "not applicable"
- CSI (consumption) questions accept number of days between (0 and 7)
- If HH has no children in school age, withdrawing children from school as a coping strategy should be not applicable
- If the HH is receiving e-voucher, food expenditure should be >= (the voucher value multiplied by number of HH members)

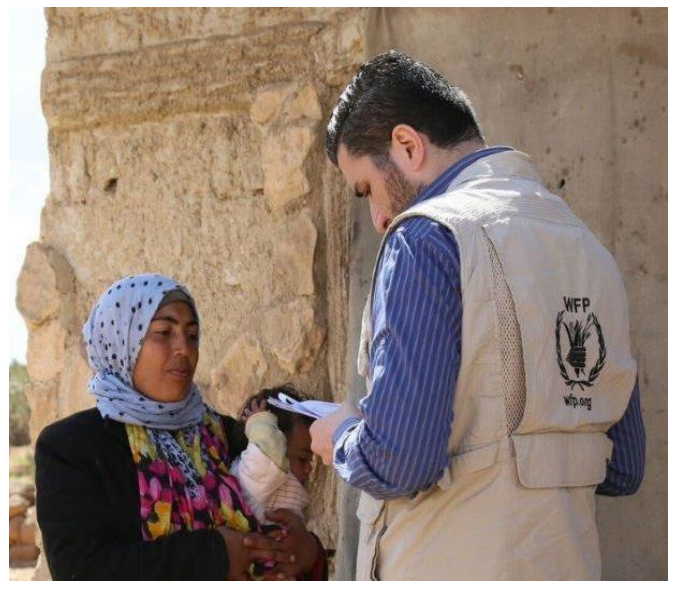
[1] <= means 'less than or equal'

Step 2: Data collection

Enumerator's role:

During the interview, enumerators should be able to validate information from respondents and probe properly.

- When HH reports low food consumption
  - Make sure that egg consumption is considered when asking about protein food consumption (egg, fish and meat)
  - In case of reporting low cereals/tubers consumption, enumerators needs to probe to make sure that the household did not consume bread everyday
- When HH reports high meat and dairy products consumption
  - Confirm the number of days of protein consumption to check if the food is consumed or is rather a condiment (e.g. minced meat used in stuffed vine leaves, or pasta with béchamel is not considered as meat consumption)
  - Confirm that food is consumed among the majority of the family in the HH (e.g. if only young child consumed milk, this not considered HH consumption)
  - Cross-check coping strategies and if dairy or meat consumption is high make sure that the number of days coping strategies were used is low



WFP Field monitor conducting face to face interview

Supervisor`s role:

1) Supervisor should review the data coming from the field on daily basis to check and take actions immediately. Checking is most important during the first day(s) and for new enumerators.

- Missing and out of range values (especially, if the data is collected manually)
- All cases of low cereal or oil consumption (<=4 days), usually it is not common to have such low consumption, check also how this low consumption is consistent with consumption of other food groups
- Very high food consumption for meat and dairy (>=6) where there is low food consumption scores for cereals (<=4)
- Check HH with high meat and dairy consumption (>=6) while reporting high coping strategy
- Check HH with very high or very low expenditure on different expenditure items
- Check for patterns of data entry like same entry repeated many times (7 , 7 , 7, 7) or alternating numbers (2,1,2,1,2,1,2,1)

2) Supervisor should call back or visit ALL household with data inconsistency in addition to a random sample of 5 -10% of HHs to check:

- If the interview took place
- Accuracy of the collected data especially the main indicators “Food Consumption” and “Coping strategies”
- Obtain feedback from the interviewee about their satisfaction with the interview

3) Supervisor should conduct training/coaching sessions with enumerators

- Based on the findings of the data quality check, individual/group coaching sessions with enumerators to brief them on the findings of the data quality checks and possible ways for improvement.
- Enumerators falling behind in the number of completed questionnaires should be contacted to identify the reasons and necessary action should be taken e.g. it may be that they are unable to operate the online questionnaire properly or they do not fully understand the questions.
- During the session supervisor should address any challenges faced by enumerators in understanding the questions.
- Periodic training should be conducted to all enumerators to insure consistency, especially if there is a high turn over.

Phone surveys

- All phone calls should be recorded (if possible)
- If recorded 5-10% of the recorded calls should be listened to for quality control
- Very short calls and very long calls should be checked

Data coming from newly joined enumerators should have a special focus in the data quality review and coaching sessions

Step 3: Data cleaning

Data cleaning takes place after finishing all data collection. **High data quality should be ensured before reaching the cleaning stage**, in order to reduce the amount of work spent on cleaning and avoid discarding data coming from non-credible enumerators. During data cleaning the whole dataset should be checked for the following:

1) Missing, out of range, and outliers entries
   - Check expenditure module for outliers and decide how to deal with odd data, sometimes HH has accidental high expenditure “marriage, health” that might skew the findings, enumerators might also include extra zeros by mistake
   - Check food consumption module for low and high food consumption of certain food groups as explained in the previous section
2) Check HH that have high positive correlation between FCS and CSI “high FCS and CSI at the same time” or “ low FCS and CSI at the same time”, usually high FCS is associated with low rCSI.
3) Data could be checked also for consistency between different enumerators
   - Calculate of the average number of consumption days for each food category disaggregated by enumerators.
   - Calculate the % of food consumption groups “poor, borderline, acceptable” disaggregated by enumerators.
   - The enumerators should have comparable results provided that they are conducting interviews in same geographical areas and the number of interviews conducted by each enumerators is big enough (>30), if one or some enumerators have outlier values, further data exploration should be followed on the interviews conducted by the identified enumerator(s). **This step is crucial if new enumerators are joining the team.**
4) Data should be checked for differences by geographical areas/sub-offices. This is crucial especially if each sub-office was trained separately.

How to identify outliers using sort in Excel

- Open the excel sheet, go to Data, then sort
- In “sort by” identify **ALL** column that should be included “e.g. food groups” or “expenditure categories”
- Specify ascending order “A-Z” or descending “Z-A” for each column
- Look at the top part of the excel sheet

How to calculate average of consumption days by enumerator in Excel

- Go to pivot tables, insert enumerators name column in the rows, in values include different consumption categories
- Change the values from “change value setting” to average

How to calculate average of consumption days on SPSS

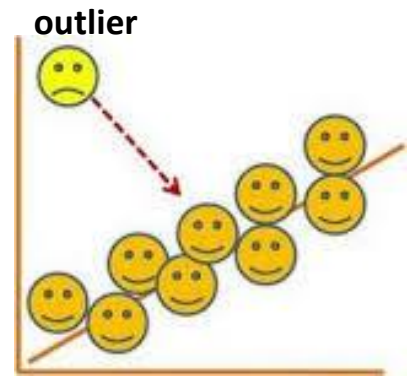
- Go to custom tables, drag enumerator in the row and different consumption categories in the column
- To get the mean calculated, make sure that the food consumption categories are identified as “scale”

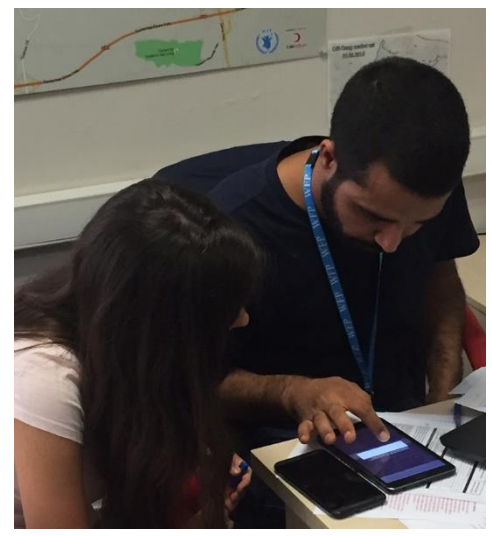
5) Decide on how to deal with outliers

- Look back into the hard copy (If it was used) or listen to the recorded calls (if it was phone survey), make sure that the questions were asked properly and the answer was recorded correctly.
- If the data was a mistake in data entry, correct it.
- If there is an outlier for expenditure due to accidental expenses, leave it in place until the analysis stage. Data analyst might decide to discard outliers later.

6) Decide on how to deal with inconsistencies

- Look back into the hard copy (If it was used) or listen to the recorded call (if it was phone survey) to make sure that the questions were asked properly and the answer was recorded properly.
- If there was a mistake from the enumerator side (e.g. the way the question was asked or no probing) check all the other records of the identified enumerator.
- If data inconsistency is noticed in several records from same enumerators, delete ALL the records of this enumerator. Individual coaching session should be conducted and panelizing action should be taken data fabrication was proved.





Field monitor conducting his first PDM through phone interview, while a supervisor is supporting his attempt, Turkey 2016

Examples from the field

Example (1) arranging the food consumption in ascending order shows that some HH had almost very low food consumption, interviews needs to be reviewed. At the data collection stage, probing showed have done with beneficiaries to identify how the HH could survive zero cereals/tubers consumption for one week (no bread, rice, potato, etc…). It could be also a data entry mistake, or error happened while converting the file!

	FC_CerealCon	FC_VegetableCon	FC_FruitCon	FC_MeatFishEggCon	FC_PulsesCon	FC_DairyCon	FC_OilFatCon	FC_SugarCon	FC_CondimentCon
1	0	0	0	0	7	0	0	0	0
2	0	0	0	0	0	1	7	0	0
3	0	0	0	3	1	2	7	1	7
4	0	0	1	1	7	0	7	7	7
5	0	2	0	2	1	3	3	0	3
6	0	7	0	3	0	0	7	0	7
7	0	7	0	3	7	7	4	7	1
8	0	7	2	7	7	7	7	0	7
9	1	0	0	0	0	0	0	0	0

Example (2) arranging expenditure in descending order shows that three HH have high expenditure under “other expenditure category”. Interviews needs to be reviewed, if the entries were correct, keep the values as it is, outliers could be discarded later at the analysis stage not to skew the results. Reporting Median expenditure instead of the mean would also help to overcome the effect of the outliers.

@6_Expenditure_and_Income_SourceG_6_2Q6_2_...	@6_Expenditure_and_Income_SourceG_6_2Q6_2_...	@6_Expenditure_and_Income_SourceG_6_2Q6_2_...	@6_Expenditure_and_Income_SourceG_6_2Q6_2_...	Other_for_comparability_with_CFSME	@6_Expenditure_and_Income_SourceQ6_expenditures
0	0	5	2000	2044	unvistety study expenses
0	0	0	1500	1549	wedding expenses
0	0	80	1000	1050	pay for other cases
0	5	0	0	380	0
0	100	0	0	365	0
0	0	0	0	358	Kerosene for heating
0	45	15	250	315	pay for other cases
0	90	32	95	259	Clothes,baby milk
0	201	0	0	231	0
0	180	30	0	230	0
350	180	22	0	210	0
0	40	5	110	208	birth expenses
0	50	25	125	205	remittances to syria
0	180	0	0	201	0

Example (3) analysis of food consumption (FC) groups by enumerators shows that two enumerator have collected **extremely different** FC results than the rest of the enumerators. In addition, the data was extremely different from other historical data collected in the same area, further review needed for the interviews done by these enumerators. Action taken to remove all the records collected by the identified enumerators.

		FCGroup							
		Poor		Borderline		Acceptable		Total	
		Count	Row N %	Count	Row N %	Count	Row N %	Count	R
1.1 Interviewer/FMA's name	bel	1	20.0%	4	80.0%	0	0.0%	5	
	Geh	7	3.0%	40	16.9%	190	80.2%	237	
	Mohar	34	41.5%	31	37.8%	17	20.7%	82	
	mohame	1	2.0%	9	18.0%	40	80.0%	50	
	Mohame	0	0.0%	3	13.6%	19	86.4%	22	
	Som	28	50.0%	22	39.3%	6	10.7%	56	

Example (4) arranging FCG and CSI in descending order, two household that have very high FCS and very high CSI, interviews to be reviewed as when FCS is very high, usually coping strategy shouldn’t be that high.

37.00

FC_CerealCon	FC_VegetableCon	FC_FruitCon	FC_MeatFishEggCon	FC_PulsesCon	FC_DairyCon	FC_OilFatCon	FC_SugarCon	FC_CondimentCon	CSI	FCS
7	7	1	7	7	7	7	7	7	10.00	106.00
7	7	1	7	7	7	7	7	7	5.00	106.00
7	7	3	7	7	7	7	2	7	.00	105.50
7	7	0	7	7	7	7	7	7	37.00	105.00
7	3	2	7	7	7	7	7	7	.00	103.00
7	7	0	7	7	7	7	3	7	.00	103.00
7	7	2	7	5	7	7	7	7	3.00	101.00
7	4	1	7	7	7	7	2	3	5.00	100.50
4	7	0	7	7	7	7	7	7	32.00	99.00

For further questions contact:
Anna-Leena Rasanen, Regional Monitoring Advisor (anna-leena.rasanen@wfp.org)
Mohamed Salem, Regional Monitoring Officer (Mohamed.salem@wfp.org)