

2019 Syria HNO Multi-Sector Assessment	Technical Process Guideline: Data Aggregation Method
Technical Note	V2

This note outlines the method to triangulate, and aggregate questionnaires collected by programmatic partners and data collection initiatives at community and neighborhood level as a part of Multi-sector assessment. Aggregation will be achieved by systematic logic outlined in the subsequent sections and implemented using R-script programming.

The following guidelines outline the process followed during the Aggregation/Triangulation phase with data collected through the Multi-sector assessment. The Aggregation/Triangulation phase immediately follows the Data Cleaning and Verification phase.

OVERVIEW

Data collected by operational partners and data collection initiatives will produce the following types of data records:

1. Neighborhood-level records
2. Community-level records

Multiple partners and data collection initiatives are collecting data in approximately 5,600 communities in Syria where majority of communities are covered by more than one partners. Therefore, many communities will have multiple community-level records for triangulation. These multiple records per community will be weighted, triangulated, and aggregated to produce one record per community or neighbourhood levels.

Neighbourhood-level records will be collected in 13 major cities in Syria where neighbourhood p-codes are available. In these cities, each neighbourhood will be considered as a unit of analysis.

The data triangulation and aggregation process are outlined below:

- 1) Step One: Recoding of data variable
The data exported from KoBo will be recoded to proper names and labels.
➤ Output: recoded RAW dataset
- 2) Step Two: Criteria and score for data confidence level
This confidence level information will be used in triangulation process.
➤ Output: data reliability criteria and scores for triangulation
- 3) Step Three: Aggregation method
For each question in the questionnaire, aggregation method will be identified and aggregation logic will be elaborated.
This process will be implemented using R-Script.
- 4) Step Four: Community/Neighborhood-Level Aggregation
The raw dataset will be aggregated at neighbourhood level records from the 13 major cities, and the community level records to produce one record per community.

- **Outputs:** One record per community or neighbourhood

Additionally, data will be aggregated to community/neighbourhood level with additional consideration of gender of community focal points (CFP).

- **Additional Outputs:** two records per community or neighbourhood (one for male CFP and another one for female CFP). Aggregated data will be separated per sector and saved in separate files.

1. Data Preparation for aggregation/triangulation

Before starting data aggregation, raw data submitted by partner will be cleaned and verified during 'data cleaning / verification' step. In addition, entries in the 'Other' are explored, analysed and processed as below.

- Individual data collected by partners are first of all merged into one raw master data base. All 'Other' entry variables are explored to identify 'Other' entries that do not fit into existing categories and have therefore not been re-coded by respective agencies. If a significant number of entries relating to a 'new' category is identified, a new variable (if binary) or new category is created and implemented across all data sets to ensure inclusion in the final aggregation of all categories. For variables where all categories are retained in the final aggregation, 'Other' will be included as a category in the event of any entry in any of the corresponding records. The text entry will remain in the raw data set for reference.

2. Criteria and Score for Confidence Level

The data weighting system is based on criteria related to the methodology applied in collecting data. The information related to the methods of data are recorded for each section/sector and will be scored as outlined in the table below. ¹

Parameters/Criteria	Options	Score
Community Focal Point (CFP) Role or knowledge in subject matter ²	Expert	3
	General	2
	Limited	1
Data Collection modality	Face to Face	3
	Remote	1

The sum of the scores for each section of the questionnaire will be used to evaluate confidence level for each questionnaire record and this composite score will be recorded in the dataset.

¹ A general framework of data weighting system for HNO 2017 was developed by "Weighting work stream" of HNO process.

² Sector to suggest score for each CFPs types

3. Aggregation Method (same geographic level)

Data Aggregation Plan

Multiple questionnaire submitted at same geographic unit will be aggregated into one record. For example, multiple questionnaires submitted by more than one programmatic partner and data collection initiative for a community will be aggregated to one record.

Following records will be generated by aggregating multiple questionnaires into one.

Type of data records	Aggregation and analysis need – Main output	Aggregation and analysis need – Additional output
Neighborhood	Neighborhood Level	Neighborhood Level for male & female CFPs
Community	Community Level	Community Level for male & female CFPs

The logic to systematically aggregate each type of question present in the Multi-sector questionnaire will be as outlined below³.

GLOSSARY

Categorical: questions and answers referring to categories that are not ordered in any way (e.g. Question: What are the main protection issues? Answers: Tension between host and displaced populations; Unexploded ordinance/Explosive devices etc.)

Ordinal: questions and answers referring to categories that are ordered in some way (e.g. Question: Damages to private buildings due to conflict? Answers: No damage; slight damage; moderate damage etc.)

Continuous/Numerical: questions and answers referring to numerical values (e.g. Question: How many IDPs are currently living in this community? Answer: e.g. 2,000)

³ Data aggregation logic is similar to the method used in multi-sector assessments for HNO 2016, HNO 2017 and HNO 2018.

Detail method description

3.1. Numerical question where 1 number is given as answer

Example Question: CCCM 2.2. How many collective centers and camps are in your community?

To estimate number closer to the questionnaires provided by the most reliable source, weighted average of the answers from different questionnaires will be calculated.

$$\text{Calculation} = \frac{\text{Data 1} \times \text{Score 1} + \text{Data 2} \times \text{Score 2} + \text{Data 3} \times \text{Score 3}}{\text{Score 1} + \text{Score 2} + \text{Score 3}}$$

Result will be rounded to nearest integer value.

Unknown values (Do not know or No answer) which are entered as -1 or -5 and converted to NA.

Questions in the KoBo where this category applies: Refer KoBo aggregation file for detail list of questions.

3.2. Numerical question where % is given as answer and SUM of categories to be 100%

Example question: 6.1 What % of your community live in each of the following type of dwelling?

To estimate the final % for each category of source following steps will be used:

- Unknown values (Do not know or No answer) which are entered as -1 or -5 and converted to NA.
- Weighted average for each category will be calculated for the figures as in 3.1
- SUM of total may exceed 100% which requires further adjustment before analysis.

Questions in the KoBo where this category applies: Refer KoBo aggregation file for detail list of questions.

3.3. Categorical question where 1 categorical answer is given

Example: 1.2b. What was the main reason for their displacement to this location? (Tick one box)

Answers from different questionnaires will be aggregated by identifying the answer provided 1) most often with 2) the highest reliability. The following steps will be followed to aggregate individual records into one record per location.

- Create binary variable for each of the categories
- Convert YES/NO or TRUE/FALSE to 1 or 0 respectively.
- Replace 1 with reliability score
- Sum the score for each category and identify the category with highest score
- Add the category with the highest score as a final result.

If multiple answers are found to have the same top score, all with the top score will be retained in the aggregated data set. Further verification and triangulated is required to establish the final result for the location.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.4. Categorical question where 3 categorical answers are given without ranking

Example: 1.6a. What are the three most recent types of assistance delivered to the community in the past three months? (Tick up to three)

Following steps will be followed to aggregate individual records into one record for the location.

- Create binary variable for each of the categories
- Convert YES/NO or TRUE/FALSE to 1 or 0 respectively.
- Replace 1 with reliability score
- Sum the score for each category to identify the categories with highest score
- Add the categories with the 3 highest score as a final result.

If multiple categories are found to have same highest score within the top 3 highest scores, all will be retained. Further verification and triangulated is required to establish the final result for the location.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.5. Categorical question where 4 categorical answers are given without ranking

Example: 2.11. What were the four biggest problems hosting and hosted households faced? (Select top four)

Following steps will be followed to aggregate individual records into one record for the location.

- Create binary variable for each of the categories
- Convert YES/NO or TRUE/FALSE to 1 or 0 respectively.
- Replace 1 with reliability score
- Sum the score for each category to identify the categories with highest score
- Add the categories with the 4 highest score as a final result.

If multiple categories are found to have same highest score within the top 5 highest scores, all will be retained. Further verification and triangulated is required to establish the final result for the location.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.6. Categorical question where 3 categorical answers are ranked 1 – 3

Example: 1.1. What are the top three concerns within this community?

(rank only top three: 1= most at risk, 2=second most at risk, 3=third most at risk)

Following steps will be followed to aggregate individual records into one record for the location.

- Scoring the rank – assign score 3 to highest rank, 2 to second rank and 1 to lowest ranked options.
- Multiply the rank score by reliability score
- Sum the score and identify the categories with the top three highest total score.
- Add these top three categories as final result together with associated rank

If multiple categories are found to have same highest score within the 3 highest

ranked scores, all will be retained. Further verification and triangulated is required to establish the final result for the location.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.7. Categorical question where 4 categorical answers are ranked 1 – 4

Example: 8.1. Which among the following groups are most at risk or vulnerable? (rank top 4) (rank top four: 1= most important, 2=second most important, 3=third most important, 4=forth most important)

Following steps will be followed to aggregate individual records into one record for the location.

- Convert rank to score – assign score 4 to highest rank and 1 to the lowest ranked options.
- Multiply the rank by reliability score
- Sum the score and identify the categories with the top three highest total score.
- Add these top 4 categories as final result

If multiple categories are found to have same highest score within the 4 highest ranked scores, all will be retained. Further verification and triangulated is required to establish the final result for the location.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.8. Categorical question where all categorical answer that apply are given

Example:

Answers from all questionnaires will be retained as the final result.

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.9. Ordinal question where 1 ordinal answer is given

Example: Education – 3.2: What percentage of children in the community are out of formal education in 2016/2017 school year?

Answers from the different questionnaires will be aggregated by giving more weight to answers provided with a higher reliability.

The following steps will be followed to aggregate individual records into one sub-district reality level record:

- Recode the ordinal categories into 1, 2, 3, 4 scores etc.
- Multiply each category score (e.g. 1, 2, 3) by the corresponding reliability score and sum up the scores

- Divide the sum of scores by the total sum of reliability score to obtain a weighted average across categories
- Recode back into corresponding original category, summing up where a value falls in the middle (e.g. 1.5)

Calculation:

$$\text{Weighted Category Score} = \frac{\text{Ordinal Category 1} \times \text{Score 1} + \text{Ordinal Category 2} \times \text{Score 2} + \text{Ordinal Category 3} \times \text{Score 3}}{\text{Score 1} + \text{Score 2} + \text{Score 3}}$$

If the weighted category score falls between 2 categories, adopt the worst case scenario.

Examples:

Education Q3.2

% range	None 0%	1 – 25%	26 – 50%	51 – 75%	76 – 99%	All 100%	Do not know	No Answer
Ordinal Score	0	1	2	3	4	5	Exclude	Exclude
Weighted Category Score (WCS) range	0 < WCS < 0.5	0.5 ≤ WCS < 1.5	1.5 ≤ WCS < 2.5	2.5 ≤ WCS < 3.5	3.5 ≤ WCS < 4.5	4.5 ≤ WCS		

Questions in the KoBo where this category applies: Refer KoBo aggregation method file for detail list of questions.

3.10. Consideration of text input in 'Other' field

- Retain all text information entered as concatenated texts.

3.11. Consideration of 'Do not know', 'Not Sure' and 'No Answer' in the aggregation

- If one of the variable is selected together with 'do not know', 'not sure' or 'no answer', then unselect 'do not know', 'not sure' or 'no answer'.
- If one of the questionnaire has response in any other main variables, consider it as final result.
- If there is no response for all main variables, return 'NA' in all variable.

3.12. Special cases

- **Health: 6b.1.1.1 Mode of Travel and minutes travelled are entered as two separate questions**
 - Mode of travel -> Car/by foot
 - Minutes travelled (for mode of travel selected)

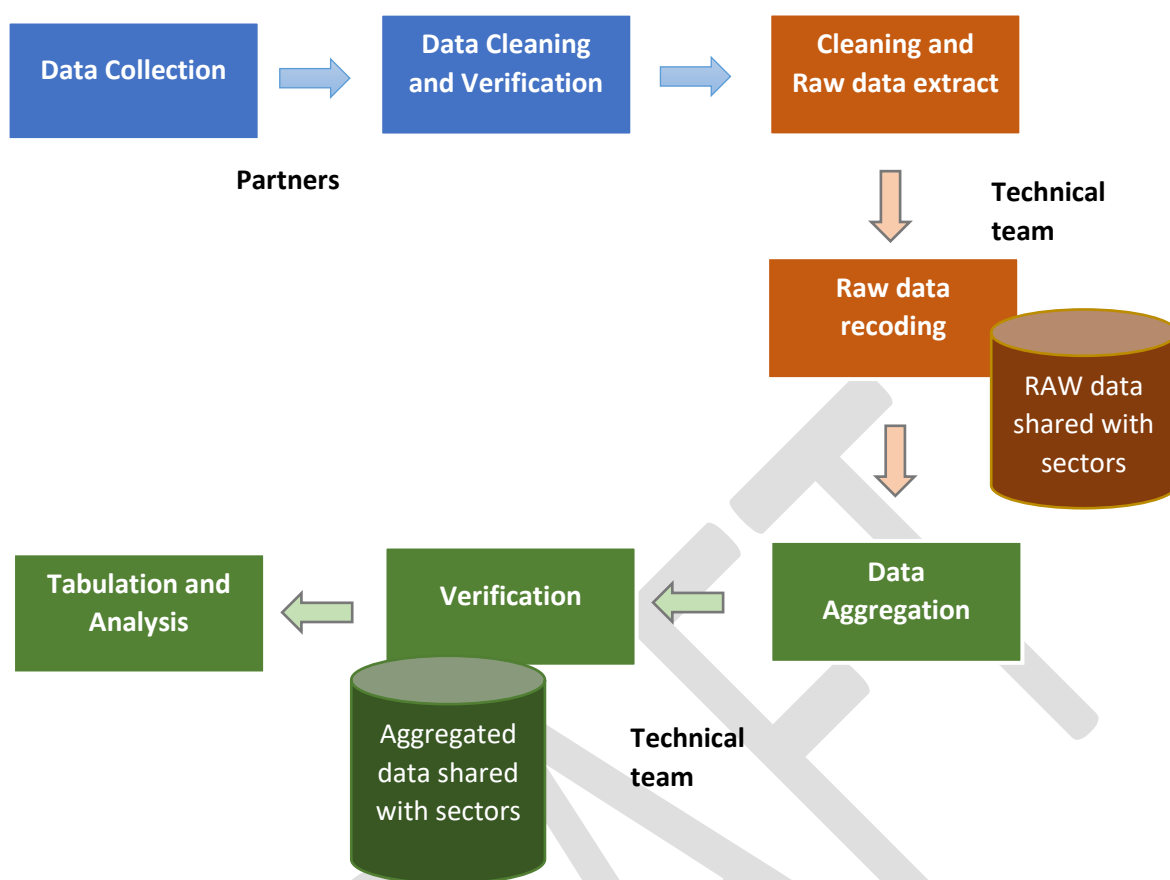
These two questions are considered as dependent questions and following process is used for aggregation.

- Add additional columns corresponding to mode of travel
- Transfer minutes travelled value to newly added respective column
- Weighted aggregation as explained in 3.1 for each mode of travel.

- **Protection: BOYS/GIRLS/MEN/WOMEN/ALL answers**
 - In case 'ALL' is selected, transfer '1' to Boys, Girls, Men and Women columns.
 - Aggregate as explained in section 3.8
 - Retains all answers.

DRAFT

Annex 1: Workflow of technical process



Annex 2: Technical process

- **Kobo xlsx form preparation**

- Aggregation methods
- Variable coding and scores

Refer KoBo aggregation method file for detail methods used.

<https://github.com/ppsapkota/kobohrtoolbox/tree/master/xlsform>

- **R-script source code**

<https://github.com/ppsapkota/kobohrtoolbox>

DRAFT

Summary of aggregation methods: multiple records to one per community

Type of question	KoBo Data Type	Aggregation Method Code ⁴	Notes on data processing	Weighting Applied	Example
Admin pcodes	select_one	ADMIN			Locations
Continuous variables where average estimation is desired	integer, double	AVG	Average	Average – no weighting applied	Age of CFP
Continuous variables where average estimation is desired	integer, double	AVG_W	Weighted Average	Confidence level weighted average	number of returnees in a location
Categorical variable where text is entered (mostly for others and metadata)	Text, Date, Select_one	CONCAT	Simple concatenation of all the responses to retain information	No weighting applied	Text in others field, gender of CFPs, Dates of data collection
Categorical variable where text is entered (mostly for metadata and locations)	Text, select_one	CONCAT_U	Simple concatenation of all the unique responses to retain information	No weighting applied	Locations
Categorical variable where single answer asked	select_one	SEL_1	Recode all TRUE response as 1 and FALSE as 0	Confidence level weighted sum and return highest ranked option	Most of the select one questions (excluding group of questions for ranking)
Categorical variable where single answer asked	select_one	SEL_1_RALL	Retain All – like select multiple answers	Retain all responses	
Categorical variable where single answer asked, and the response is related to a category selected in another question	select_one and related to integer question	AVG_W_SEL_1_REL	Response related to data entered in another question. Weighted Average	Confidence level weighted average	First, mode of transport selected (e.g by car, by foot) and then minutes required is entered.
Categorical variable where up to three responses asked	select_multiple	SEL_3	Recode all TRUE response as 1 and FALSE as 0	Confidence level weighted sum and return top three ranked options	

⁴ Method code is used in R-Scrip for different type of aggregation.

Type of question	KoBo Data Type	Aggregation Method Code ⁴	Notes on data processing	Weighting Applied	Example
Categorical variable where up to four responses asked	select_multiple	SEL_4	Recode all TRUE response as 1 and FALSE as 0	Confidence level weighted sum and return top four ranked options	
Categorical variable where multiple responses asked	select_multiple	SEL_ALL	Recode all TRUE response as 1 and FALSE as 0	Retain all responses	
Categorical variable where responses are asked to RANK	select_one	RANK3 or RANK4	Recode all rank to corresponding score such as 3 for first rank and 2 for second rank etc.	Confidence level weighted sum of score and return top three ranked options and retain rank order. (Borda count mean method)	
Ordinal categorical variable	select_one	ORD_1	Assign score to the ordinal variable and process based on assigned score.	Confidence level weighted average of score and return variable corresponding to the average score. In case of score falling in the middle, adopt worst case scenario.	
Ordinal categorical variable	select_one	ORD_1_RUP	Assign score to the ordinal variable and process based on assigned score. Round up the score.	Confidence level weighted average of score and return variable corresponding to the average score.	Mostly used for YES/NO confirmation questions.
Ordinal categorical variable	select_one	ORD_1_WCASE	Assign score to the ordinal variable and process based on assigned score	adopt worst case scenario	Used for alternate optional output for the protection cluster.

Type of question	KoBo Data Type	Aggregation Method Code ⁴	Notes on data processing	Weighting Applied	Example
Metadata for confidence level	select_one	SCORE	convert to score and present average score of multiple data records	Average of metadata score	modality of data collection, Type of CFP
All questions related to additional information or not relevant for higher admin level aggregation		NA			

DRAFT