

# VAF 2024 METADATA

## RESEARCHER'S GUIDE

### METHODOLOGY BRIEF

2024 VAF used a stratified random sampling approach to select cases with a margin of error below 5 per cent at 95% confidence level. Stratification was planned along two variables: nationality (Syrian, Iraqi and Other) and location (Hosting communities and Camps). Syrians were divided into subgroups for each of the twelve governorates. While Syrian refugee households were grouped depending on their governorate of residence, non-Syrians were divided between Iraqi and Non-Syrian/Non-Iraqi with two geographic units assigned for each: Amman or other governorates (North, Central, and South). The governorates were grouped into North, Central and South to avoid sampling less than 50 non-Syrian cases in certain governorates. After grouping Syrians, Iraqis and non-Syrians into these geographic units, a random sampling and an oversampling strategy was used to select cases.

This sample was randomly drawn from cases registered in the ProGres registration database administered by UNHCR Jordan. The host community sample includes refugees residing in urban, peri-urban and rural settings and excludes those living in refugee camps. The camp sample have been selected using a different sampling frame. For camp sampling, the two largest refugee camps, Azraq and Zaatari, which host only Syrian refugees, were selected.

### DATA STRUCTURE

While the sampling design was based on “cases” or UNHCR registration groups, the design of the data collection tool also allowed households, sharing groups, families, and individuals to be introduced as distinct grouping levels in the research. The following lines clarify the distinctions between the different units of analysis:

- 1- **Individuals:** refers to any single individual living in the household that have been targeted for the survey regardless of type of relationship they have with the focal point of the selected family. However, it is important to note that for some VAF analysis only registered refugees' data who were living in the same household of the selected family were considered for the reporting.
- 2- **Cases:** The term “case” generally refers to UNHCR's registration groups, which consists of a refugee or an asylum-seeker and their dependents who are treated as a single unit for the purpose of services and assistance. This grouping typically includes a principal applicant along with their family members, who are dependent on or make up a part of the principal applicant's household. Cases are used for assessing eligibility and needs, processing applications, and providing protection and aid. However, this typical grouping mechanism may not accurately represent the full spectrum of household configurations, particularly in instances where the familial structure extends beyond the nuclear model to include additional kinship ties, as is often observed in extended family settings. This necessitates looking at different levels of individuals' grouping units.
- 3- **Family:** This term represents a nuclear family, typically two parents and their children, most times corresponding to UNHCR's registration group. Some households can be composed of several families (i.e multiple cases), in rare instances multiple cases could represent a single family (e.g. if each parent is in a different registration group).

- 4- **Sharing groups:** a group of individuals who share a dwelling and share meals and expenses with other groups is named a “sharing group”. This unit was introduced in the 2022 VAF to better understand how refugees are living together in Jordan: whether several families live together and share resources, or they occupy the same dwelling to cut on rent expenditure, while not sharing resources. This allows to administer modules such as consumption and expenditure at the sharing group level for greater accuracy.
- 5- **Households:** a group of related or unrelated individuals who share the same dwelling (residing together and living under the same roof), irrespective of their pooling of resources or resource-sharing arrangements.

Accordingly, in the microdata a **tab that corresponds to each level of data collection can be found**, with two exceptions:

- Caravans in camps: The “caravans” is a section with a dedicated loop to ask certain questions about the standardized shelters in camps commonly known as “caravans” – it is important to not that some questions were camp-specific such as the dimension of caravan which has been specific to Zaatari Camp.
- Sharing Groups having a short and a long version which are put in separate tabs. The details of each group can be found in their respective XLSForm.

While modules such as income and cash assistance were administered at the case level, health, education, and livelihoods were administered at the individual level. Factors associated with multi-case households were believed to provide useful insight into additional dimensions of vulnerability that could be missing if cases are treated only as separate entities: indicators such as monthly rental payment or food consumed were also easier for a respondent to recall at the household or sharing group level as these resources are commonly shared.

## LONG VS. SHORT FORMS

A household consumption and expenditure module have been deployed at the sharing group level for all households participating in the survey. To enhance response rates while avoiding survey fatigue, a two-version approach was employed using a long and a short version of the questionnaire.

Both short and long survey forms, cover non-food items (e.g., cleaning materials, clothing, education, health, housing, personal care, transport, tobacco, and durables) and food items (e.g., fruits, grains, proteins, beverages, and restaurant meals). The total number of items is 62 for the short form and 96 for the long form.

The most significant items consumed by the poorest non-Jordanians (bottom 20%) were identified from the Household Income and Expenditure Survey (HIES). These items were ranked, and the top items from each sub-category were selected based on their share of total food consumption and the number of households consuming them. Two consumption modules were created: a short form with 62 items and a long form with 96 items. 75 percent of the sample in each location completed the short form, and the remaining 25 percent completed the long form. The long form covers items that account for about 90% of total consumption by the poorest non-Jordanians.

Food consumption data were collected using a recall method, asking households about their consumption in the seven days before the interview. The survey was administered in two steps: first, determining if any household member consumed, purchased, or received food items as aid; second, asking follow-up questions about the quantity consumed or received. The long form also included questions about the cost and quantity of purchased items. Enumerators were trained to help identify quantities.

Non-food consumption had different recall periods depending on the sub-category, ranging from one month to annually. The consumption modules for camp and out-of-camp populations were similar, except for durable items included only in the camp module. These durable items had a negligible impact on total consumption.

## NAVIGATING THE DATA EXPORT

The uploaded Excel file contains the following sheets:

### 01\_HOUSEHOLD

This sheet contains data related to household characteristics. This may include information on household facilities, and housing-related details. The first 12 columns in this sheet are household and head-of-household identifiers.

### 02\_SHARING LONG

This sheet focus on consumption expenditure data through the standard expenditure module. This is composed of three sub-modules in their extended version:

- Food submodule (seven-day recall)
- Non-food submodule (30-day recall)
- Non-food submodule (six-month recall).

Additionally, this sheet covers livelihood coping strategies adopted. The first 14 columns in this sheet are household and head-of-household identifiers plus the sharing group size and index.

### 03\_SHARING SHORT

This sheet focus on consumption expenditure data through the standard expenditure module. This is composed of three sub-modules in their short version:

- Food submodule (seven-day recall)
- Non-food submodule (30-day recall)
- Non-food submodule (six-month recall).

Additionally, this sheet covers livelihood coping strategies adopted. The first 14 columns in this sheet are household and head-of-household identifiers plus the sharing group size and index.

### 04\_FAMILY

This sheet provides detailed data on family units within households. There were no specific questions on this level it has been mainly used to identify nuclear families within the household. The first 16 columns in this sheet are household and head-of-household identifiers, sharing group index and size, plus the family size and index.

### 05\_CASE

This detail cases within the dataset, this sheet provide data related to income and assistance as well as other case-related information. The first 18 columns in this sheet are household and head-of-household identifiers, sharing group index and size, the family size and index, plus the case size and index.

## 06\_INDIVIDUAL

This contains individual-level data. It includes detailed personal information on all individuals in the dataset, such as demographics, employment, education, health, and other socio-economic factors. The first 19 columns in this sheet are household and head-of-household identifiers, sharing group index and size, the family size and index, the case size and index, plus the individual index.

## 07\_WEIGHT

This sheet contains weighting variables used to ensure that the dataset accurately represents the population. It includes weight values for different groups or categories, which can be applied during analysis to account for survey design or sampling methods. Weights are also included in each tab for easier reference.

### MERGING DATA AT DIFFERENT LEVELS:

#### DATA IDENTIFIERS:

As mentioned above, the data comes from two questionnaires that have different (sharing group level) variables, accordingly it would be possible to have duplicate index for surveys that used the long sharing group module with the surveys that used the short sharing group module.

To create a unique identifier, we need to create a composite index by concatenating the existing indices with the "Form" column. This step is crucial because indices across different forms might have identical values. By generating a composite index, we ensure uniqueness across all forms.

The composite index will be used to trace data through the hierarchy. Using the variable Form with any of the following indices:

- *Householdindex*
- *Sharingindex*
- *Familyindex*
- *Caseindex*
- *individualindex*

For each record, concatenate the index with the "Form" column. For example, if the index is '12' and the form is 'Long', the composite index should be '12Long'. Similarly, for a 'Short' form, it should be '12Short'. Apply this step consistently across all levels to maintain uniformity. Always use the composite indices as the key for merging data.

#### An Example:

To merge individual level data with the household level data, at the individual level we will merge the householdindex with the form, also at the household level we will merge the householdindex with form. Matching indices from both data sheets are collected from the same household.

## INTRODUCTION TO DATA COLLECTION TOOL

This section is dedicated to elaborating on the data collection tool for individuals who do not have background on the tool.

KoboToolbox is an open-source suite of tools designed for data collection in challenging environments. It is widely used by researchers, humanitarian organizations, and various other groups for its ability to handle data collection efficiently and accurately, even in remote or disaster-affected areas. The primary components of KoboToolbox include an easy-to-use web-based form builder, robust data storage and management systems, and flexible data analysis and visualization tools.

### How KoboToolbox Works:

**Form Creation:** The process begins with creating a form using the KoboToolbox form (XLSForm). Users can include various types of questions such as text, multiple choice, numerical inputs, and media uploads.

**Data Collection:** Once the form is created, it can be deployed to data collectors. KoboToolbox supports both online and offline data collection through its mobile app, KoboCollect.

**Data Management:** Collected data is stored securely in the KoboToolbox server, where it can be accessed for analysis.

### XML Code vs. Label in KoboToolbox

**XML Code (variable name):** This is the underlying structure of the form, which provides the variable name that refers to the questions, for example age is the variable name that refers to the questions “what is your age?”, so instead of showing the entire question you would have a short name that refers to the question. Which is optimal for most data analysis software and database structures that may not allow variable names to exceed a certain threshold or restricts certain characters.

**Label:** Labels are the user-friendly names and descriptions assigned to questions and choices in a form (i.e. the full question). They are what the end-users (enumerators) see when filling out the form. Labels make the form understandable and accessible without requiring knowledge of the underlying XML structure. While labels are easier to work with, they offer less flexibility for advanced customization compared to XML code.

It is important to understand that data exports are provided in xml codes and can be referenced to the label using the XLSForm, the details of the process is detailed in the following section.

## READING A KOBO XLSFORM

Kobo XLSForm is a standardized form format that uses Excel to create and manage survey forms. Understanding how to read and interpret an XLSForm is crucial for effectively designing and deploying surveys using KoboToolbox. The form typically consists of multiple sheets, with the most important being survey and choices.

---

## IMPORTANT COLUMNS IN KOBO XLSFORM

### 1. Type:

The type column specifies the type of question or input expected. It defines what kind of data will be collected. Common types include:

text: For open-ended text responses.

integer: For numeric responses.

select\_one [list\_name]: For single-choice questions from a predefined list.

select\_multiple [list\_name]: For multiple-choice questions from a predefined list.

### 2. Name:

The name column defines the variable name for each question. This name is used internally to identify the data field and should be unique within the form. It must follow naming conventions, typically using lowercase letters, numbers, and underscores without spaces.

### 3. Label:

The label column provides the text that will be displayed to the respondent. It is the user-friendly question or instruction that appears on the form. Labels can be customized for different languages if multilingual support is needed.

### 4. Choices:

The choices sheet is linked to the survey sheet through list names specified in select\_one or select\_multiple question types. It includes the following key columns:

list\_name: Groups of choices for specific questions.

name: The unique identifier for each choice option, used in data storage.

label: The text displayed to respondents for each choice option.

Example:

| type              | name          | label                |
|-------------------|---------------|----------------------|
| text              | respondent_id | Respondent ID        |
| select_one gender | gender        | What is your gender? |
| integer           | age           | How old are you?     |

**Choices Sheet:** This sheet lists the choices for select questions.

**list\_name:** Groups choices under a common identifier.

**name:** Unique identifier for each choice option.

**label:** The text displayed for each choice option.

Example:

| list_name | name   | label  |
|-----------|--------|--------|
| gender    | male   | Male   |
| gender    | female | Female |

## ONE-HOT ENCODING IN KOBOTOOLBOX

**One-hot encoding** is a technique used to transform categorical data into a format that can be provided to machine learning algorithms to do a better job in prediction. In KoboToolbox, this can be particularly useful for preparing survey data, especially for multiple selection questions, for advanced analysis such as statistical modeling or machine learning.

One-hot encoding converts categorical variables into a series of binary columns, each representing a unique category within the original variable. Instead of having a single column with multiple categories, you end up with multiple columns (one for each category), and for each observation, only one of these columns will be marked with a 1 (indicating the presence of that category), while the others will be marked with 0.

## XLSFORM AS A DATA DICTIONARY

As detailed above, in XLSForm, the 'survey' sheet contains the questions and their types, while the 'choices' sheet includes the options for multiple-choice questions. Each question is defined with a 'name' (the variable name used in the dataset), 'type' (the question type such as text, integer, select\_one), and 'label' (the question text displayed to users). Additional columns can specify constraints, default values, and relevance conditions. This structured format allows for clear documentation of each variable's purpose and properties, facilitating data collection and analysis. By organizing survey elements in this way, XLSForm serves as a comprehensive data dictionary that ensures consistency and clarity throughout the survey lifecycle.

XLSForm can be found in the micro-data library downloads tab.

## SAMPLING

### SAMPLING STRATEGY

#### INTRODUCTION

The VAF 2023 aimed for a total sample size of 8,653 households (HHs) to be interviewed in both urban and camp areas. The sample divided as follows:

- 5,100 Cases for Syrian Out of Camp
- 2,000 Cases for Non-Syrian Out of Camp, comprising:
  - o 1,200 Iraqis stratified across Amman vs. non-Amman
  - o 800 refugees of other nationalities stratified across Amman vs. non-Amman
- 1,563 Cases for Syrian in Camp

#### SAMPLING PROCEDURE

The sampling procedure considered the **population size proportionally, with the proportion varying in between strata**. This proportion has been driven by **the number of cases** not individuals since it is a household survey -. This helped ensure larger samples to cover larger populations. However, some stratification decisions are influenced by context and analysis approaches. This includes the decision to sample Syrian refugees, Iraqis, and non-Syrian non-Iraqi refugees as separate groups. This choice reflects the fact that Syrians represent almost 80% of refugee population and Iraqis make up nearly half of the non-Syrian refugee population, making them the second-largest nationality group among the registered refugees.

#### DATA UNITS

Despite data units here being cases (Asylum Seeker Certificates or Proof of Registration documents), which typically represent a family registered on the ProGres database as active refugees, the actual survey targets households. This means that if a sampled case lives in the same household as another sampled case, this will result in a single visit to that household and consider this a one step towards the target not two.

##### 1. Syrian Out-of-Camp Refugees

For Syrian out-of-camp refugees, the stratification procedure remains consistent with that of VAF 2021. Syrian refugee households are grouped by their governorate of residence. A random sampling strategy was used to select cases to interview in each governorate, with oversampling to ensure necessary margins of error within strata. The targeted sample proportion dedicated to Syrian refugee HHs in urban areas is 5,100, equivalent to approximately 60% of the total sample size of 8,653.

##### 2. Non-Syrian Out-of-Camp Refugees

For non-Syrian out-of-camp refugees, the stratification is based on the refugee's nationality (Iraqi or Non-Syrian/Non-Iraqi), with two geographic units assigned for each: Amman or Other governorates (North, Central, South). A random sampling and oversampling strategy was used to select cases to interview. The targeted proportion dedicated to non-Syrian refugee HHs in urban areas is 2,000, equivalent to approximately 23% of the total sample size of 8,653.

### 3. Camp Sampling

For camp sampling, the two largest refugee camps, Azraq and Zaatari, hosting only Syrian refugees, were selected. The selection of the number of refugee households was based on the probability proportional to size criteria. The targeted proportion for Syrian refugees in camps is 1,563, representing approximately 18% of the total cases sample size of 8,653. Then this was distributed amongst Azraq and Zaatari proportionate to size.

#### STRATIFICATION UNITS

The 2023 sampling methodology continued delineating between Iraqi and refugees of other nationalities (Non-Syrian, Non-Iraqi) with a modification in the stratification per governorate.

This results in an overall stratification by Country of Origin: Syrian (12 governorates), Iraqi (2 units), Non-Syrian Non-Iraqi (2 units).

##### Syrian (Urban)

| Strata     | Region  | Population<br>(# cases in September-2023) | Sample<br>Target |
|------------|---------|---|------------------|
| Al Balqa   | Central | 4,743                                     | 300              |
| Amman      | Central | 59,585                                    | 1,100            |
| Madaba     | Central | 3,381                                     | 300              |
| Zarqa      | Central | 13,270                                    | 450              |
| Ajloun     | North   | 1,517                                     | 300              |
| Al Mafraq  | North   | 21,420                                    | 550              |
| Irbid      | North   | 36,320                                    | 700              |
| Jarash     | North   | 2,343                                     | 300              |
| Al Aqaba   | South   | 1,112                                     | 300              |
| Al Karak   | South   | 2,287                                     | 300              |
| Al Tafilah | South   | 367                                       | 200              |
| Ma'an      | South   | 2,095                                     | 300              |
| Total      |         | 148,440                                   | 5,100            |

##### Iraqi (Urban)

| Strata | Region              | Population<br>(# cases in September-2023) | Sample Target |
|--------|---------------------|---|---------------|
| Unit 1 | Amman               | 23,302                                    | 700           |
| Unit 2 | Central/North/South | 2,905                                     | 500           |
| Total  |                     | 26,207                                    | 1,200         |

##### Non-Syr Non-Iraqi (Urban)

| Strata | Region              | Population<br>(# cases in September-2023) | Sample Target |
|--------|---------------------|---|---------------|
| Unit 1 | Amman               | 9,432                                     | 450           |
| Unit 2 | Central/North/South | 2,350                                     | 350           |
| Total  |                     | 11,782                                    | 800           |

#### Syrians (in-camp )

| Strata        | Region   | Population<br>(# cases in September-2023) | Sample Target |
|---------------|----------|---|---------------|
| <b>Mafraq</b> | Za'atari | 16,726                                    | <b>1,004</b>  |
| <b>Zarqa</b>  | Azraq    | 9,157                                     | <b>549</b>    |
| <b>Total</b>  |          | <b>25,883</b>                             | <b>1,553</b>  |

This approach has several advantages:

- Comparability with the VAF 2021 data across the three nationality groups: Syrian, Iraqi, Non-Syrian/Non-Iraqi.
- Ensures that at each stratum we maintain a 5% margin of error at 95% confidence level.
- Avoids sample sizes with fewer than 50 cases in certain governorates for non-Syrians which could yield data that could not be generalized.
- Ensures a sample size of more than 350 households per stratum for Non-Syrians
- For Iraqi cases, it combines governorates with fewer than 60 cases
- For Non-Syrian/Non-Iraqi Cases, it combines governorates with fewer than 200 cases.

## SAMPLE WEIGHTS

Weights in sampling are crucial for ensuring that the sample accurately represents the population. When certain segments of a population are overrepresented or underrepresented in a sample, weighting adjusts for these imbalances. This process assigns a weight to each observation, reflecting its relative importance or frequency in the overall population. Incorporating weighting into analysis helps to correct for sampling bias, improves the precision of estimates, and ensures that conclusions drawn from the sample data are generalizable to the broader population. Without proper weighting, the analysis could be skewed, leading to inaccurate or misleading results.

| Location  | Nationality    | Weight     | Achieved Sample Size | Population (#Cases) | Population Percentage |
|-----------|----------------|------------|----------------------|---------------------|-----------------------|
| Azraq     | Syrian refugee | 15.6596155 | 653                  | 9157                | 4.31%                 |
| Zaatari   | Syrian refugee | 14.835679  | 1259                 | 16726               | 7.88%                 |
| Ajloun    | Syrian refugee | 4.1519737  | 310                  | 1517                | 0.71%                 |
| AlBalqa   | Syrian refugee | 12.194666  | 330                  | 4743                | 2.23%                 |
| AlKarak   | Syrian refugee | 6.4680834  | 300                  | 2287                | 1.08%                 |
| Amman     | Syrian refugee | 38.9487    | 1298                 | 59585               | 28.06%                |
| Aqaba     | Syrian refugee | 3.4060867  | 277                  | 1112                | 0.52%                 |
| Irbid     | Syrian refugee | 30.330729  | 1016                 | 36320               | 17.11%                |
| Jerash    | Syrian refugee | 6.6044483  | 301                  | 2343                | 1.10%                 |
| Maan      | Syrian refugee | 5.6429243  | 315                  | 2095                | 0.99%                 |
| Madaba    | Syrian refugee | 9.467455   | 303                  | 3381                | 1.59%                 |
| Mafraq    | Syrian refugee | 32.166344  | 565                  | 21420               | 10.09%                |
| Tafila    | Syrian refugee | 1.9221256  | 162                  | 367                 | 0.17%                 |
| Zarqa     | Syrian refugee | 19.752714  | 570                  | 13270               | 6.25%                 |
| Amman     | Iraqi refugee  | 59.193966  | 334                  | 23302               | 10.98%                |
| Non-Amman | Iraqi refugee  | 9.407529   | 262                  | 2905                | 1.37%                 |
| Amman     | Other refugee  | 21.004364  | 381                  | 9432                | 4.44%                 |
| Non-Amman | Other refugee  | 9.632261   | 207                  | 2350                | 1.11%                 |

In a typical household survey, weights may be composed of several factors: selection, non-response, and post-stratification. Best practices generally recommend accounting for attrition as well as selection when constructing weights. However, in cases where the sampling frame is equal to the whole population, post-stratification may not be necessary. This careful construction and application of weights ensure that the survey results accurately reflect the entire population, leading to more valid and reliable conclusions.

The weights in the urban dataset are calculated using the sample weights and non-response. In camps, the non-response was similar across the camps, and the sampling was proportional to the population, so weights are not included in this data set. For conducting joint analysis across the whole refugee population, the analysis would need to weight to allow for the relative sample proportions.

---

## USING SAMPLING WEIGHTS:

### Calculate Weighted Average:

To calculate a weighted average, divide the weighted sum by the total sum of weights.

THE FORMULA FOR THE WEIGHTED AVERAGE ( $W\_AVG$ ) IS:

$$W\_AVG = (\sum (X\_i * W\_i)) / (\sum W\_i)$$

where:

$x\_i$  is the value of observation  $i$ .

$w\_i$  is the weight of observation  $i$ .

$\Sigma$  denotes summation.

### Calculate Weighted Frequency:

For each unique categorical item, count the number of occurrences of that item in the dataset, considering the sample weights.

THE FORMULA FOR CALCULATING THE WEIGHTED FREQUENCY ( $W\_FREQ$ ) OF A CATEGORICAL ITEM IS:

$$W\_FREQ\_I = \sum (W\_i * I(X\_i = ITEM))$$

where:

$w\_i$  is the weight of observation  $i$ .

$x\_i$  is the categorical item of observation  $i$ .

$I(\text{condition})$  is an indicator function that equals 1 if the condition is true and 0 otherwise.

$\Sigma$  denotes summation over all observations.

THE FORMULA FOR CALCULATING THE WEIGHTED PERCENTAGE ( $W\_PERC$ ) OF A CATEGORICAL ITEM IS:

$$W\_PERC = (W\_FREQ\_I / (\sum W\_i)) * 100$$

## OUTLIER DETECTION

In this study, outliers were detected and managed using a **Winsorization technique**. Specifically, the lower and upper bounds were set at the 1st and 99th percentiles of the data, respectively. Any value that fell below half of the lower bound or exceeded twice the upper bound was considered an outlier. To mitigate the impact of these outliers, they were replaced with the median value of the dataset. This approach ensures that extreme values, which could skew the results, are controlled without reducing the sample size. Winsorization helps maintain the integrity of the data while providing robustness against the influence of anomalous values.

Specifically, the methodology involved the following steps:

**Setting Lower and Upper Bounds:** The lower bound was set at the 1st percentile, and the upper bound was set at the 99th percentile of the data distribution.

**Finding the Median Value:** The median value of the dataset was calculated, serving as a measure of central tendency that is robust against extreme values.

**Identifying Outliers:** Values were identified as outliers if they were below half of the lower bound or above twice the upper bound. These thresholds were chosen to capture extreme values that significantly deviate from the majority of the data.

**Replacing (Imputing) Outliers with the Median:** Outliers identified in the previous step were replaced with the median value of the dataset. This replacement ensures that extreme values do not disproportionately influence the results while preserving the overall data structure.

## KEY LIMITATIONS

**Sampling among UNHCR-registered refugees:** The sample was drawn randomly from UNHCR's ProGres registration database. Consequently, it is only comprised of cases that have maintained their status as registered refugees with UNHCR. The VAF has always excluded refugees who have never been registered with UNHCR (unless living within a targeted household). As a result, the results of the survey may not accurately represent vulnerabilities of the refugee population that is not registered with UNHCR.

**Respondent bias:** The methodology relies on self-reported levels of a household's socio-economic situation. As with any form of self-reporting, there is potential for inaccuracies and bias. There is also a risk of bias associated with the (perceived) power differences between the enumerator and the respondent, as some cases may have responded to survey questions with the aim of demonstrating their eligibility to receive assistance or other services in the future. To minimize the impact of this bias, enumerators were trained in providing comprehensive counselling on the purpose of the interview, obtaining informed consent, and conducting referrals to relevant UNHCR units as and when they were required.

**Sensitive and protection-related information:** VAF is a household survey, and the interview is conducted with the head of household, or other adult household member. Obtaining accurate information on sensitive areas related to protection risks (gender-based violence, child abuse, etc.) is not always possible in this context, and such questions were intentionally omitted with the understanding that other approaches are more appropriate to capture sensitive topics. Enumerators were however trained to recognize a potential protection concern, and a separate and secure protection referral form was used to inform UNHCR Protection teams.

## KEY VARIABLES AND DEFINITIONS

- Governorate: The administrative division where the household is located.
- District: The smaller administrative division within a governorate where the household is located.
- CaseNo: A unique identifier assigned to each survey case representing the head of household for the sampled case.
- householdindex: An index or identifier specific to each household in each survey short or long. Should be combined with "Form" to be unique across the survey.
- Form: The type or version of the survey form used for data collection (Long or Short).
- sharingindex: An index or identifier specific to each unit of shared living arrangements within a household in each survey short or long. Should be combined with "Form" to be unique across the survey.
- familyindex: An index or identifier specific to each family unit within a household. Should be combined with "Form" to be unique across the survey.
- caseindex: An index or identifier for each individual case within the household, family, or sharing unit. Should be combined with "Form" to be unique across the survey.
- CaseNationality: The nationality of the household is identified as the nationality of the head of household.
- HoHID: Head of Household Identifier; a unique identifier for the head of the household.
- householdsiz: The total number of individuals living in the household.
- sharingsiz: The total number of individuals sharing a living arrangement within the household.
- familysiz: The total number of individuals within a family unit in the household.
- casesiz: The number of individual cases surveyed within a household, family, or sharing unit.
- nationality: The nationality of household identified as the nationality of the focal point (main person on the UNHCR registration document) for each household identified into the three main strata (Syrian, Iraqi, and Other).
- gov: Abbreviation for Governorate; the administrative division or province of the individual.
- Syrian: A binary variable indicating whether household is headed by a Syrian refugee.
- weight: The survey weight assigned to the household, used for statistical analysis to adjust for sampling design.
- individualindex: An index or identifier specific to each individual within the survey short or long. Should be combined with "Form" to be unique across the survey.
- Enumerator\_name\_ind: a unique ID of the enumerator who conducted the survey.

## XLSFORM EXPORT

As detailed above, the XLSForm can be used as a data dictionary. The link to the files can be found here:

Host Communities: <https://microdata.unhcr.org/index.php/catalog/1030>

Camps: <https://microdata.unhcr.org/index.php/catalog/1098/>