

IDH Farmfit

Farmer Survey

Primary Data Collection

Methodology

Latest Update: January 2023

Table of contents

Table of contents	2
Data Collection Methodology	5
1. Background and Rationale	5
2. Objective	5
Who are these guidelines for?	6
How was the methodology for primary farm-level data collection created?	6
Principles guiding design?	6
3. PDC Content: What data?	8
About Farmfit Primary Data Collection Question Library	8
Core questions	8
Case questions	10
Optional questions	10
4. PDC Process: How to collect data?	13
1. Design	13
1.1 Intake	13
1.2 Survey Adjustment	14
Structure of the Farmfit PDC Questions Library	14
How to use the question library?	16
Data Quality in Survey Design	18
1.3 Sampling	18
1.4 Field Plan (Data collection plan)	19
1.5 Enumerator Selection	20
2. Capture	20
2.1 Enumerator training	20
2.2 Data collection	21

Data collection during COVID	21
2.3 Real time data quality control	22
3. Understand	23
3.1 Data cleaning	23
3.2 Data Delivery	24
5. Timeline and teams	25
6. Ethical Considerations	26
7. Guiding Principle for Updates to Methodology	27
ANNEXES	28
Annex 1: Data Anonymization Protocol	29
1. Anonymisation of data for Farmfit Primary Data Collections	29
2. Joint Controller Data Sharing Protocol	30
Annex 2: Enumerator Field Guide	32
Annex 3: Policy document: Monitoring with enumerators during COVID 19	36

This document was produced by Akvo (the data collector) upon request of the IDH Farmfit team. While the aim is for the Primary Data Collection Methodology to be a standardized process, learning is iterative and, thus, this document and its accompanying tools will continue to evolve so that methods continue to be fit-for-purpose.

Data Collection Methodology

1. Background and Rationale

Farmfit is a program funded by Melinda and Bill Gates Foundation (BMGF) and the UK's Foreign, Commonwealth and Development Office (FCDO) implemented by IDH, that seeks to transform inclusive agricultural markets. Its target is to increase the income of 1.1 million smallholder farmers by at least 30%, in five years.

Farmfit assumes a market based approach: companies that offer services to farmers - such as those selling fertilizers or purchasing crops for export - request Farmfit/IDH to help them analyze and improve the services they offer to smallholder farmers. IDH has developed a standardized analysis tool to determine the efficiency, effectiveness, sustainability and scalability of smallholder engagement models, which is referred to as Service Delivery Model (SDM). The analysis is done by working with the companies to gather 1) financial company data on the services delivered to smallholder farmers and 2) farm economics and other characteristics of their producer base.

For the second component, the farm economics and other characteristics of the producer base, accurate information on smallholder farmers is needed. From July to December 2019 a standardized approach for collecting primary data from smallholder farmers was developed. The approach was revised in 2020 to finetune the methodology and integrate relevant components that derived from the review of:

- Farmfit's Learning Framework
- Farmfit's Impact modules
- Research on actual household income measurement to benchmark against living income
- IDH's farmer segmentation tool

2. Objective

The objectives of this document are to provide guidance on how to implement primary data collection at farming household level to meet the needs and standards of the IDH Farmfit Program.

Who are these guidelines for?

These guidelines are created with two audiences in mind:

1. Akvo staff working to collect farm level data for IDH's Farmfit Program (IDH has established a framework agreement with Akvo for primary data collection)
2. External data collectors seeking to align with data collection standards used within the Farmfit Program

How was the methodology for primary farm-level data collection created?

The methodology, primarily reliant on a household survey, was initially designed jointly with the IDH Farmfit and M&E team. Research themes guiding the design of the household survey were refined during a design workshop at the start of the pilot (2019) and took into account the Farmfit program's Theory of Change, monitoring framework, the farm-level data needs for the SDM analysis and the insights from key IDH Farmfit staff. The original and revised versions of the survey and methodology took into account review of existing instruments and methods including:

- [Feed the Future Monitoring, evaluation and Learning Toolbox](#)
- [Living Standard Measurement Study- Integrated Surveys in Agriculture](#) (World Bank)
- [Rhomis](#)
- [LICOP guidance](#)

IDH also shared documents, offered guidance and feedback to various versions of the surveys and to this document.

Principles guiding design?

The key principle guiding the development of this methodology is a balance between **practicality** and **rigour**. The aim is to maintain the burden on farmers to a minimum while generating quality data that responds to IDH Farmfit's needs. Given the nature of the Farmfit program's work across commodities, countries and service delivery models, the methodology also balances **standardization** and **modularity**, as a means to scale and remain cost-effective. In practice this means that there are elements of the data collection process that have been standardized and some elements that need to be adjusted per each data collection case (modularity concept). **A case** refers to the request for primary data collection related to a company that the Farmfit team works with and where an SDM analysis is being conducted. A case has geographical, crop and service delivery boundaries.

An overview of the Standardization and Modularity principles, in practice, is presented in the table below.

Standardization



- Question library with core modules
- Intake form
- Training methodology
- Enumerator guidelines
- Monitoring dashboards
- Data cleaning
- Data delivery

Modularity



Case specific:

- Permits/approvals
- Survey adaptation (i.e. optional modules)
- Sampling
- Data collection plan

To be able to deliver within budget and timeframe, the data collector and IDH agreed on the following boundaries to the scope of data collection:

Scope of data collection		
Factor	Scope	Implication
Population definition	Smallholder farmers ¹	Medium and large scale farmers that are associated with the company are not taken into account in SDM analysis.
Sample size	Up to 375 farmers per case	Stratification of farmers is out of scope for standard data collection that is in line with the bronze evaluation standard.
Geography	1 to 2 areas (e.g. provinces, districts...)	SDM company which operates nation-wide needs to choose 1 or 2 geographic areas that are representative of their portfolio.
SDM Crop	1 to 2 SDM crops	SDM company which focuses on various SDM crops needs to choose 1 or 2 crops to focus on.
Reference period	The last 12 months	The data reflects a 12 month reference period, which is prone to recall bias.

Unit of analysis

The unit of analysis for the Farmfit PDC survey is the individual farmer. The survey contains questions related to the household and farm that are addressed by the person who is in charge of the farm. The section related to gender calls for the female head of household to respond to the questions.

¹ Smallholders are defined by the SDM companies, as the definition of smallholder varies per country and crop.

Similarly, the food security questions are addressed to the household member who generally prepares the food for the household.

3. PDC Content: What data?

The objectives of primary data collection at farm level for the Farmfit program are:

1. To better understand characteristics of farmer households and provide descriptive and numerical data as inputs for the SDM analysis
2. Generate data for Farmfit indicators (to be used for baseline and endline measurement)
3. Generate comparable aggregate data that allows comparison and learning across companies/SDMs

About Farmfit Primary Data Collection Question Library

Data needs have been consolidated from various sources into a comprehensive **Farmfit Primary Data Collection (PDC) Question Library**. The Farmfit PDC question library is an Excel database of survey questions that have been formulated to respond to data needs identified by the IDH Farmfit team.

The question library was created upon IDH Farmfit team request to support data collectors standardize the content of the Primary Data Collection survey by ensuring the same questions are used consistently in the same way. Having a comprehensive set of questions can also reduce the time invested in designing of data collection instruments that aim to meet the same purpose.

The question library can be received upon request.

As introduced in the previous sections, there are elements of the data collection process that have been standardized and some elements that need to be adjusted per case. Practically, this means that the question library consists of core, case-specific or optional questions. Case-specific and optional questions can be selected or "tagged" in the PDC question library to be added to the Core PDC survey.

Core questions

Core questions are those that need to be included in all PDCs for Farmfit. These questions respond to data needs that are essential for SDM analysis and MEL purposes. Core questions fall into the following Core categories:

Core Categories²	Content description (examples of content below are non-exhaustive)
Household Characteristics	Characteristics of the household members including the number and gender of household members, the gender of the head of household and Food security status (Months of Adequate Household Food Provision - MAHPF)
Farmer Profile	Characteristics of the lead farmer in the household, including the person's age, sex, educational level, land ownership, possession of mobile phone or mobile banking account. Also includes perception questions related to their intent to continue to farm or whether farming remains their best investment.
Farm Characteristics	Characteristics that are inherent to the farm, not related to farming practices. This includes the farm location, the land size, water sufficiency (optional), or crop specific characteristics related to perennial crops.
Farm Practices	Activities and tasks performed by the farmer(s) in the farm. Farming characteristics not inherent to the physical farm. This includes volumes of crop produced, sold, lost, or used for own consumption. Also includes variables related to farm labour, equipment and input use, and crop protection mechanisms.
Farmer Services	Services farmers receive such as training, membership services (coop/associations), premiums/certifications, access to financial services, customer satisfaction, contracting, payment behaviours.
Farm economics	Variables necessary to estimate crop income and net farm income.
Metadata	Metadata is defined as "a set of data that describes and gives information about other data." Metadata provides context to the data files that result from each data collection case so users may place the data in time and place. Includes variables such as date and time of data collection, individual survey duration time, and unique identifier per entry (i.e. Farmer). Akvo Flow automatically generates meta data for each case, where each submission has a unique identifier with associated metadata.

All primary data collection exercises intended to meet IDH Farmfit Standard MUST contain all questions and modules labelled as Core in the PDC Question Library. The core questions are linked to Farmfit's theory of change, impact modules and learning framework and respectively correspond to the bronze level of the Farmfit evaluation framework.

² Note that categories included in this table are not the same as sections in the survey. Categories are meant to reflect the data needs that underpin Core elements of the Farmfit PDC question library. Questions are organized in a different order and sections in the survey, as the order of questions take survey design best practices into account.

Case questions

Case questions are questions that are only applicable in specific cases because of their specific *country* and *type of crop(s)*. The PDC Question library stores these questions so they may be available in the event a future data collection occurs in a country or about a crop type already been collected. Additional country- and crop- specific questions will be reviewed and permanently added to the question library every quarter (see Section 7. Guiding Principle for Updates to Methodology for more information on the process to update the question library).

- **Countries:** The country tag identifies questions that are specific to a country. The questions are labelled with the name of the country where they should be applied. An example of this tag are questions related to the Poverty Probability Index (PPI), an asset measurement index that is optional within the Farmfit methodology and which requires country-specific sets of questions. The same questions are applied per country, despite the crop or service type.
- **Type of crops:** This tag identifies questions that are specific to a type of crop. At the moment this guide is written, the library contains specific questions developed for Farmfit PDC cases focusing on *perennial crops*, *a crop that is produced in different types/varieties*, *cocoa* or *coffee*. A crop can belong to one or more of these cases. By implementing more Farmfit PDC cases in the coming years, more case specific questions for other types of crops will be added.

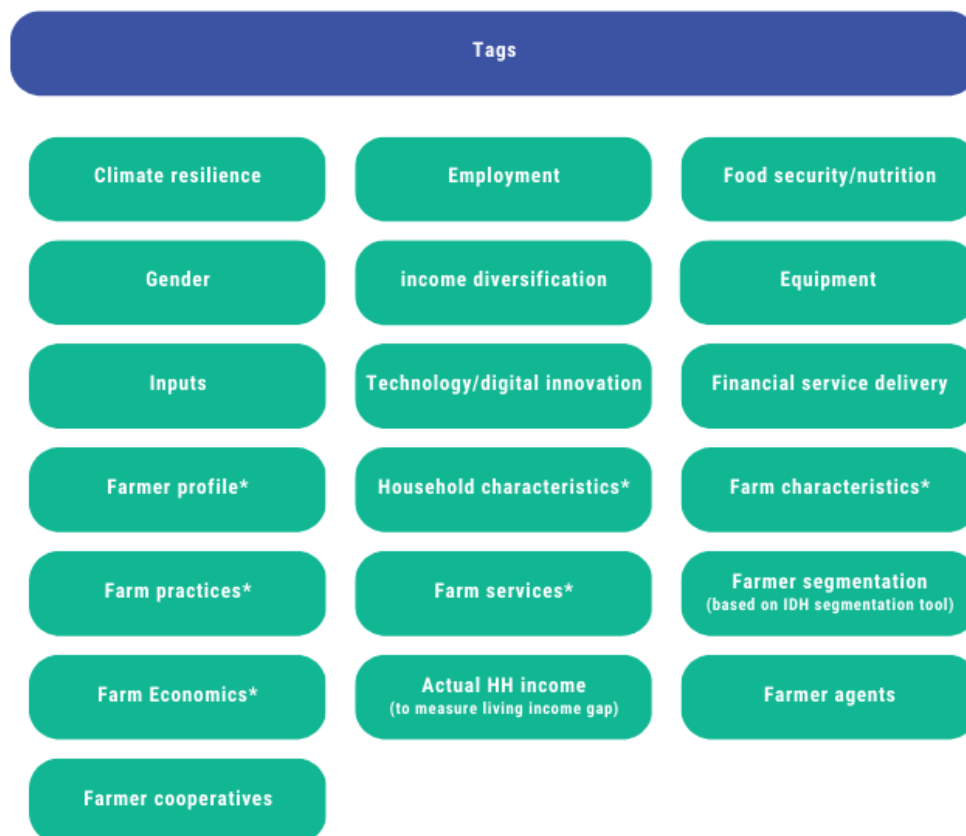
Optional questions

Optional questions are linked to thematic “tags” and allow for modular adjustments of the survey. This is referred to as a “deep dive” analysis on a specific theme. If a data collection case merits a deep dive into a particular topic, for example, gender, that case is identified as a deep dive case by the IDH staff and the company.

A deep dive case contains all core survey questions and has additional questions related to the theme of interest. All of the themes captured by the deep dives are explored at a basic level using the core survey only, but can be assessed more extensively by asking all questions that are part of that deep dive. Following the gender example, core questions allow for disaggregation by the gender of the farmer and ask the female head of household about her participation in productive and reproductive activities and decision making in the household. A gender deep dive requires the gender composition of all household members and educational levels for each. Overall, a deep dive refers to the need to have a more thorough understanding of a theme, which requires additional questions and may have sampling implications depending on the inferences that need to be derived from analysis (i.e. the need to say something about female farmers that is representative of the population, would require a stratified sample, therefore increasing the sample size).

Once a case is selected, the data collection team assigned to the Farmfit case will contact the SDM company to gather general information about the case. This is done during a kick-off meeting by using the **Intake form**. The data collection team and the SDM company can together decide which deep dives and hence which optional questions to include. The PDC Question Library allows users to select the modules that need to be included in a survey depending on the specific data needs. Sections [1.1](#) and [1.2](#) of this document outline the intake and survey adjustment processes, respectively.

In the question library, users can use the following tags to filter questions according to themes, and thereafter select the questions they choose to explore more in depth:



Tag definitions (tags with asterisk in the table above were described in the previous section as they also correspond to the core categories of the survey)

Climate resilience: Questions explore types of extreme weather events faced, duration of extreme weather effects by type, crop loss, and crop loss prevention mechanisms adopted.

Employment: This tag includes questions about on and off farm labour, including number of people hired, days hired, wage, employment type (permanent versus casual), gender or labour hired,

Food security and Nutrition: Questions capture own consumption of focus crop in the household, the MAHFP, food preference availability, water access, perception of water quality, and access to sanitation.

Gender: Questions with this tag allow for data disaggregation by capturing the sex of the farmer, the head of household, and household members. Questions also capture women participation in off farm enterprise activities (casual vs permanent by enterprise type), unpaid household labour, decision-making over productive and reproductive household activities. The survey calls for the female head of household to respond to questions regarding decision making in the household.

Income diversification: Explores first and second highest income crops including types, yield, revenue, own consumption, and loss. Also includes questions about livestock income, farm size, non farm labour net income, and non farm non labour income.

Equipment: Explores equipment used on-farm by type, costs, ownership and the source (where farmers buy equipment). The type of equipment asked about is dependent on contextual and crop specific information gathered on the intake form.

Inputs: Includes inputs used on farm, by type, costs, and challenges purchasing inputs. The type of input asked about is dependent on contextual and crop specific information gathered on the intake form.

Technology/digital innovation: Includes questions about ownership of mobile phone and functionalities, mobile money account, internet, and data use for farm activities.

Financial service delivery: Includes access to bank account, mobile money account or microfinance. Explores access to loans, including whether a farmer has received one, the source, the purpose, the amount, number of payback months, total costs, interest rates. Also explores farmers' cashflow, access to savings account, amount, reason for saving, and coping mechanisms for financial stress.

Farmer segmentation: Draws on IDH's new segmentation tool to capture questions related to farm size, on farm livestock income, market type (where she sells), who farmers sell to, contract to sell, contract duration, access to technology such as internet and mobile phone (ownership and

functionality), data use for farm activities, future outlook, water availability/sufficiency, household demographics (age, sex, education level), and farmer ID and contact details.

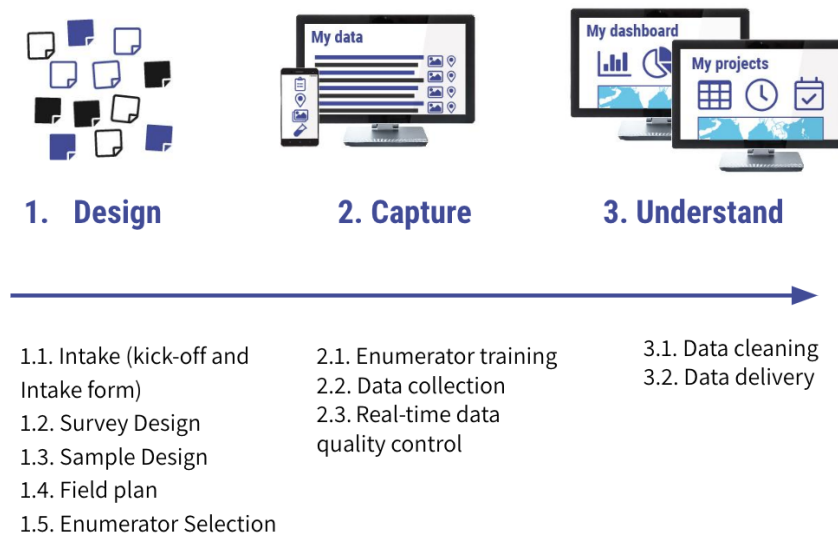
Actual HH income: Includes all questions necessary to estimate actual household income. This differs from the economic model tag in that actual HH income explores in more detail revenue from livestock and other farm crops and non-farm non income.

Farmer agents: Explores the presence of farmer agents, their main purpose and contribution, how satisfied farmers are with the agent relationship, and the amount of support an agent is providing.

Farmer cooperatives: Questions on how important farmer cooperatives are, the services they deliver, how contracts are put together, and the level of satisfaction from the individual farmer with the cooperative.

4. PDC Process: How to collect data?

The process is organized into three main phases: Design, Capture, & Understand. Main activities under each phase are:



1. Design

1.1 Intake

Once an SDM case is approved for primary data collection, the first step is for the data collection team assigned to the case to gather general information about the case from the company. This is done during a **kick-off meeting** by using the **Intake form**, previously referred to as the information needs form. The Intake form can be found as a tab in the Farmfit PDC Question Library.

The Intake form is meant to gather information from the company's local team of agronomists that is needed to:

1. Ensure compliance with the scope criteria agreed upon by IDH and the data collector, see scope in section [Principles Guiding the design](#).
2. Adjust the PDC survey to local context - for example, the information is used to define answer options for multiple choice survey questions contributing to data quality and minimizing outliers.
3. Share with the SDM company the content of the core survey and learn about specific data needs they may have. Given the comprehensiveness of the core survey, and taking survey length into account, it is not expected that companies will need to add questions to the survey other than those related to potential deep dives. In the event that a company has additional data needs, these will be discussed with the IDH Farmfit and data collector team to decide it's addition to the survey.
4. Define the sampling approach and estimate sample size,
5. Plan logistics for data collection.
6. Gather insight in reasonable values for certain variables. For example, the intake form provides ranges on variables such as farmgate prices and other data that is used to validate the primary data against contextual information.

The intake form is pre-filled by the SDM company staff and then discussed and verified during initial conversations with the IDH team and data collector. The form is then discussed and validated with the lead agronomist or head of field operations of the SDM company with the data collector. In some cases, additional input is received from a local agronomic expert on the specific crop/country context to ensure the survey is tailored to the specific situation. Akvo is leading the procurement of this expert, while IDH staff will facilitate connections when possible.

1.2 Survey Adjustment

Once the intake form is complete, the question library is used to identify the final set of questions relevant for the data collection. The sections that follow explain, first, the structure of the question library, followed by guidance on how to use it to make survey adjustments.

Structure of the Farmfit PDC Questions Library

The Survey library contains a long list of questions that can be used to create surveys. It is split into four broad sections, each with subsections described below (color-referenced): **Question selection** (colored in red), **Questions** (colored in Green), **Tags - including case & deep dives** (colored in blue), and **Functional variable description**.

Question library part	Column title	Purpose and use	Filter option
Question selection	Selection of farmfit PDC questions	Indicates if a question is part of the survey or not. All <i>core questions</i> are automatically included and cannot be deleted. If you want to include an <i>optional or case question</i> in the survey, select “In survey” in this column.	By selecting only the cells that contain “in survey”, you will see the complete survey. This will include both the core, case and optional questions you have decided to include.
	Core/ optional/ case	Indicates the use of each question as explained in section About the Farmfit PDC Question Library .	Filter questions by their use: core, optional, or case.
Survey questions	Question text	Lists the survey modules and survey questions belonging to each module	
	Answer options general	Lists the answer options and/or format for each question - excluding the questions for which the answer options are case specific.	
	<i>3 columns:</i> Answer options for specific crop	List the answer options and/or format for questions where the answer options are dependent on the focus crop. For those questions, the “general answer options” are not indicated or not applicable.	
	Question help for enumerator	Lists a “question help” for each question to be used by the enumerator when surveying a farmer. It supports the enumerator to identify the response that needs to be entered or to provide a broader explanation on the question to the farmer if required.	
	Question adjustment required by data collection team	For questions that might need to be adjusted to a specific context, this column indicates what needs to be adjusted. There are different types of adjustments, including changes in the phrasing of the question itself or creating relevant answer options using the intake form. When using the survey, make sure that all relevant adjustments are made for each question included in the survey.	
TAGS: Cases & Deep dives	Country	This column lists which questions are only applicable for one or more specific countries.	Use the filter button to identify if there are specific questions for the country where the Farmfit PDC case will take place.

	Types of crops	These four columns identify questions that are applicable for specific types of crops. At the moment these guidelines are issued this includes: perennial crops (general), cocoa, coffee, or crops produced in different varieties/types.	Use the filter button to identify if there are questions that are required to include for the crop of interest.
	<i>Thematic tags</i>	Tags questions that belong to a certain theme or deep dive. The deep dives were listed in section Optional Questions . Note that a question can belong to more than one theme and therefore showcase several tags.	Use the filter button to identify what questions belong to a certain theme.
Functional Variable Descriptions	Question use	This column identifies the function that each question has within the survey. This includes whether the response to the question is meant to be used for <i>reporting</i> of data, as a procedure question such as to support skip logic, for sampling purposes, or to support adjustment of questions that must be adapted to context (e.g. measurement unit). The intention of this column is to: 1) Support analysis of data by clearly identifying questions that need to be used for reporting (vs those that are meant to facilitate running the survey such as skip logic), 2) To inform users of why the question is included.	Use the filter button to identify the use of each question.
	Topic & sub-topic	These two columns identify the topic a question belongs to and supports the users in learning what topics are addressed in the survey.	Use the filter button to identify the topics addressed in the survey.
	Intake form	For those rows that describe elements captured of the intake form, this column indicates the section number of the intake form to which this topic belongs. The intention of this column is to link the content of the intake form to the question library.	
	Variable name	Lists the variable name related to each question, used for data delivery.	

How to use the question library?

The basic Excel filter function allows users to view specific rows in the question library, while hiding the other rows. The Excel filter is added to the second row of the question library. A drop-down menu appears in each cell of the header row by using the filter button in the second header row. Practically, this means that the question library allows you to view the questions belonging to a specific category

in line with those explained in the previous sections. For example, the filter button allows you to view all core questions or questions belonging to a specific case or deepdive.

Column A of the Question Library indicates what questions are in the survey and which ones are not. All core questions are by default marked as "In Survey." To include optional questions into the survey, the user MUST change the status on column A to "In Survey." The first column can be used to view what the survey looks like by filtering the rows marked as "In Survey".

There are three main types of adjustments required on the question library to come up with a final set of questions for each case:

1. *Filtering* to include all relevant modules: Filter through the question library tags to include all modules that are relevant for the Farmfit PDC case:
 - a. Core questions: All Farmfit PDC cases must contain all questions labelled as core.
 - b. Case questions: Use the *case filters* to assess which case questions should be included based on the context of the Farmfit PDC case (country and crop).
 - c. Optional questions: Use the *deep dive filters* to assess which optional questions should be included based on the interest of the SDM company to explore a specific thematic area.

To include an optional or case question in the survey, select "In survey" in the first column.
2. *Contextualizing*: Once you have the full set of questions needed to meet the specified data needs for the case, use the information gathered on the intake form to adjust the questions to the relevant context. The column titled "Question adjustment required by data collection team" of the question library identifies the questions that require adjustment and indicates what type of adjustment is required based on information from the intake form or the context of interest.
3. *Ordering of questions*: In general, a survey should start with less intrusive questions, such as those related to the farm, and move towards more personal or sensitive topics through the middle and end with general questions. While the question library is organized with this general arc in mind, once all optional questions are added, it is important to always check the order of questions before applying the survey in the field, so as to reduce any biases that may derive based on the order of the questions.

Akvo Flow users: The full Farmfit PDC Question Library is available on Akvo Flow. The three steps mentioned above can be done directly on the survey editor by making a copy of the master survey, deleting modules that are not relevant for that specific primary data collection case (step 1 above)

and then proceeding with steps 2 and 3 directly on the survey editor. The order of the survey questions in the question library are aligned with the order of survey questions in Akvo Flow. Therefore, we recommend you to not change the order of survey questions in Excel before editing the survey in Flow.

Data Quality in Survey Design

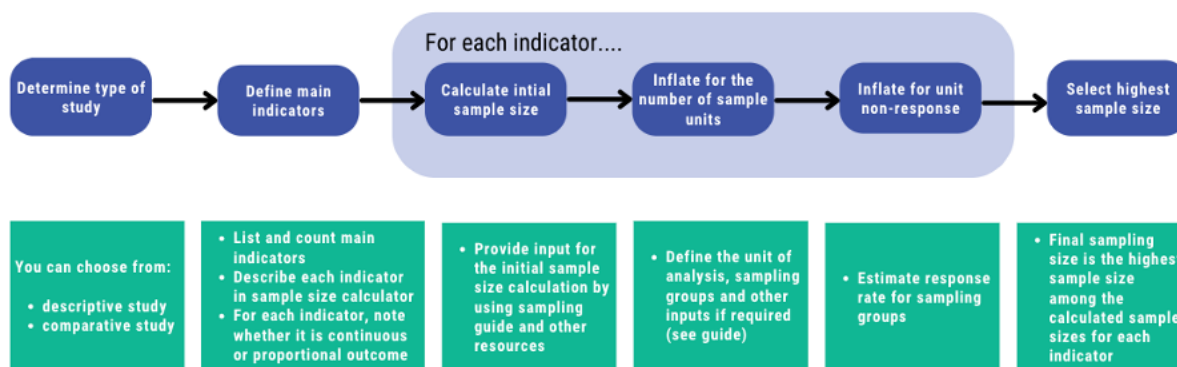
Built-in software features of the mobile data collection tool Akvo Flow, like skip-logic, mandatory questions, double entering of numerical values and adding minimum and maximum values to numerical questions, prevent the enumerator from making mistakes and reduce the risk of inconsistencies in the data. Exhaustive answer options, for example “I don’t know” and “I prefer not to answer this question” are added as answer options for multiple choice questions to minimize empty values or incorrect data.

Furthermore, open text questions are avoided as much as possible, as it leads to ambiguity and inefficiency. As each case is highly contextual, answer options may not always be exhaustive. Therefore, the option ‘other’ including a free text field is added for most multiple choice answers to not lose important information. These answers are cleaned with standard text cleaning, such as removing punctuation and setting all uppercase words or letters to lowercase, in order to recognise recurring answers.

Additional data quality measures are taken in training enumerators to interpret questions and facilitate data collection effectively with farmers.

1.3 Sampling

Primary data collection for Farmfit SDM analysis is based on power-based sampling. This approach was chosen based on the Farmfit team's need to compare results at different points in time in order to assess performance on key indicators. Power based sampling takes into account an estimate of the expected changes of key indicators, therefore providing greater confidence that the sample size will capture the expected effect at the desired level of significance. The procedure of this technique is visualized in the figure below.



Akvo has developed specific guidelines and a sample size calculator to determine sample size. The Sampling Guide and calculator can be accessed upon request along with this guide and the Farmfit PDC question library.

Please note that sampling may vary from the generic approach described in the guide and calculator based on the types of inferences that want to be drawn from the data. In the event, for example, that a deep dive case is expected, sampling may consider stratification so as to generate data that is representative of a specific population subgroup. Furthermore, the sampling approach defined is not adequate for impact evaluations where a counterfactual is expected.

In some cases, very little is known about the population that is surveyed, and in this case, applying power-based sampling is not feasible. When this happens, the population-based sampling is used as an alternative. This technique determines a sample size based on the total population size and a significance level. Another alternative sampling strategy is the snowballing or referral method, in this case almost no information on farmers is available. More information on the used sampling method can be found in the data delivery report that comes along with the data delivery.

1.4 Field Plan (Data collection plan)

Before sending enumerators to the farmers, a data collection plan is created detailing which enumerators visit which farmers and on which days. Creating a data collection plan remotely without knowing the context / geographic area is difficult. Therefore, the data collection plan is usually designed together with a representative from the SDM company and/or local extension officers before the start of the training workshop. This is already discussed and agreed upon with the SDM company during the kick-off with Akvo. Additionally, it is recommended that this is mentioned in the Farmfit business development process to ensure the company is available to provide the guidance during the data collection.

The following factors were considered while designing the data collection plan:

- Is there an equal distribution in travel time between enumerators?
- An extra margin of farmers is needed for each day, in case farmers are not home.
- Farmers need to be informed about the interview at least one day in advance.
- Formation of enumerator groups, in case they need each other's help.
- Field guides for each enumerator group to help them locate selected farmers

1.5 Enumerator Selection

The number of enumerators recruited for each case depended on the sample size and the time available for data collection (usually 5 to 6). Each enumerator on average collects five data points a day. On average the survey takes 45 minutes per farmer, but enumerators need enough time to reach the next interviewee and conditions in the field can be difficult. The following was used for enumerator recruitments:

Number of data points needed / (5 days * 5 data points) = number of enumerators

If the SDM company has worked with enumerators before, these enumerators are contacted, because their knowledge of the area and the local farmers supports a smooth data collection process. If the SDM company has no prior experience with enumerators, a ToR is shared at the nearest university that offers a specialization in agriculture. This ensures recruitment of well-educated enumerators with enough agricultural knowledge to understand the survey well.

The following factors are considered while selecting enumerators:

- Prior experience with mobile based data collection
- Background in agriculture
- General understanding of basic financial concepts (i.e. profit/loss)
- Fluency in local language/dialect
- Familiarity with the geographic area of data collection

2. Capture

2.1 Enumerator training

Once enumerators are identified, the data collector facilitates a one and a half day training. The training workshop consists of learning how to use the data collection application on the smartphone, understanding the survey, practicing interviewing techniques, and learning how to troubleshoot

during field data collection. A training session for data collection related to the Farmfit generally includes:

Planning	Module	Description
0,5 day	Mobile data collection	<ol style="list-style-type: none"> 1. Downloading mobile data collection application 2. Getting to know all the features of the application 3. Calibrating GPS signal on the phone 4. Battery saving options in the field
0,5 day	Understanding the survey	<ol style="list-style-type: none"> 1. Background of primary data collection 2. Going through survey question by question 3. Survey best practices and informed consent
0,5 day	Practicing (in the field)	<ol style="list-style-type: none"> 1. Practicing the survey in the field or amongst each other (if the field is too far from training location).

At the end of the training workshop, the enumerators have group discussions on challenges encountered in the field and have the opportunity to clarify questions about the survey. An [enumerator field guide \(annex 2\)](#) is available, which explains agricultural and financial concepts that enumerators can fall back on during data collection if they have difficulties with answering specific questions.

2.2 Data collection

During data collection, enumerators follow the field plan and guide to collect data from farmers.

Data collection during COVID

At the time these guidelines are written, the world is still facing the global COVID-19 pandemic. In light of this situation and in the spirit of continuing data collection where possible, Akvo issued a policy document that outlines basic principles for monitoring activities during the pandemic. See the document in Annex 3.

In addition to the policy, Akvo's technical consultant team developed new software features to facilitate data collection via phone. A pilot was conducted in early 2020 and results indicated the effectiveness of the approach in getting people to respond to the survey. In contexts where in-person data collection is not advisable, Akvo has adapted its software and training to conduct surveys via phone.

A key limiting factor of the phone-based interview model is that a database of phone numbers needs to be present or easily accessible (from a local company). Furthermore, it only works if there are no

questions where the enumerator needs to independently observe behaviour or an asset. The current IDH FarmFit Core survey does not strictly require that. Short surveys (to reduce phone fatigue) and monetary incentives (for example phone credits) are success factors. Also introductory calls to set a date and time for the interview and explaining to the farmers the purpose of the interview helps to increase participation.

2.3 Real time data quality control

During data collection, incoming data is visualised near real-time in a data collection tracking dashboard. This allows data collection supervisors to get an immediate overview of the GPS location of data collection, survey time, as well as the number of data points collected by each enumerator. Furthermore, error-prone questions are visualised in plots (for example: acres of SDM crop vs. kilograms of SDM crop produced) to scout outliers. In case of doubts of data quality, the supervisor calls enumerators to ask them for an explanation and/or to clarify the question.

Apart from the data quality dashboard, Akvo’s data science team performs spot checks during the data collection, in order to review data quality. Conducting a thorough review of the data quality by looking at question dependencies, common enumerator mistakes and questions that have not been answered, mistakes are detected early on. This allows the data collection supervisor to mitigate data quality issues.

The charts below showcase a section of the Data collection dashboards that are used to track data collection.



Example of the data collection tracking dashboard

3. Understand

3.1 Data cleaning

Different measures are taken to arrive at a clean data set and minimise errors. First of all, outliers are removed from the dataset. To determine outliers for the numerical questions of the survey, a cut off of three standard deviations from the corresponding mean is used. Higher or lower values than this cut off, are set to '9997' and not incorporated in further analysis. Furthermore, open text answers are cleaned with standard text cleaning, such as removing punctuation and setting all uppercase words or letters to lowercase, in order to recognise recurring answers.

Measurement units are converted to a single measurement after the data is collected. By default, in the data delivery, all land size measurements must be converted to acres and all crop measurements to kilograms to ensure this aligns with metrics and units used as part of the Farmfit SDM database.

Additional variables are created calculating the different cost components, revenues, and the resulting actual income and productivity. Also, a Net Promoter Score is deviated from the collected data.

This procedure can be automated to a large extent using R, and R markdown. R markdown incorporates R code with text in a clean and orderly way. This makes it easy for analysts to reuse the code while making it easy for people who do not use R to read and understand the transformations. A standardised R script is run by the data collector to perform all data cleaning steps (see annex 4). Next to the standardised procedure, each case also includes a manual data quality check of data related to income and productivity. Outliers and odd numbers are discussed with the data collection supervisor, and the involved enumerator to find out the reason for deviation, and try to check if we can still correct the value. Examples of such cases are when an enumerator accidentally puts too many zeros at the end, or reports the total revenues from a crop when we ask for the price. In other cases, the outlier might still be a realistic value and should be kept in the dataset. If not, the outlier is put to NA. This manual check is the main reason why a data cleaning and delivery for a case cannot be done in a very short time frame, since the sense-making of the outliers could be time-consuming when enumerators need to be contacted.

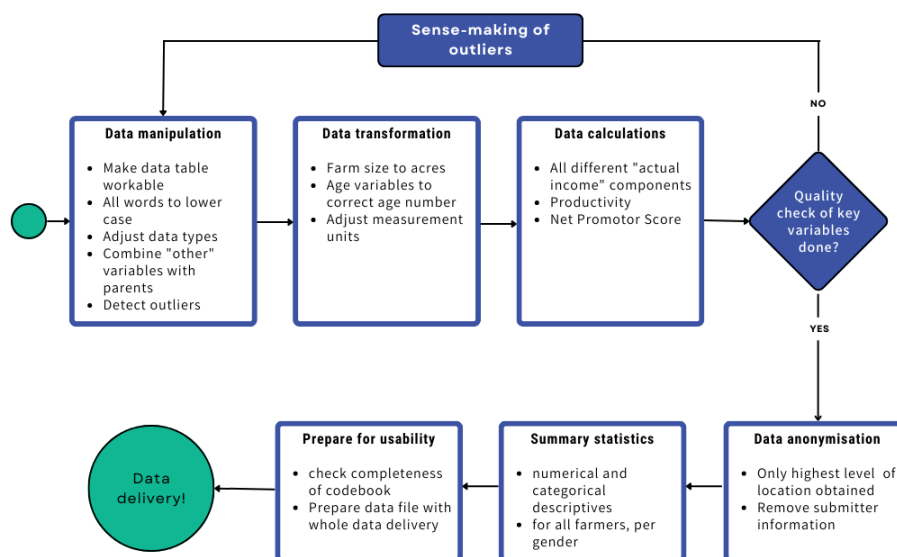
After these checks, the data is anonymised. For further use of the data, it is important that private information is deleted. Only the highest level of the farmer's location is retained. Also, the information about the enumerator is deleted.

Finally, some basic statistics are calculated for the survey variables, which allow for quick visualisations of the survey result. For numerical questions the minimum value, maximum value, the average and the standard deviation are determined. The standard deviation is calculated by taking the square root of the sample variance, which we determine with the following formula (Stats package, R core team):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

For categorical questions the frequency of the options is determined and compared to the number of farmers that filled out the question by means of a percentage. This is done once for all farmers, for each gender separately, and for other disaggregation based on the case's needs.

This data cleaning procedure is visualised in the figure below.



3.2 Data Delivery

Data is analysed using R, and R markdown is used for the delivery of the indicators to the IDH analysts. The data delivery consists of an excel sheet containing a cleaned data file and descriptive statistics of a set of indicators depending on the case, a report repeating information on the general cleaning steps and information about the sample size, and a handover meeting to discuss the main highlights of the concerning case.

1. Data delivery excel sheet with following tabs:

- Codebook
- Cleaned data
- Raw data
- Numerical descriptives farmers

- e. Categorical single descriptives farmers
 - f. Categorical multiple descriptives farmers
 - g. Numerical descriptives by gender
 - h. Categorical single descriptives by gender
 - i. Categorical multiple descriptives by gender
 - j. Net promoter score
2. **Calculation sheet:** This sheet explains the rationale behind the calculated variables.
 3. **Data delivery report:** This report explains briefly what the sample characteristics are, and what important data cleaning steps are performed to arrive at the data delivery. Further, notes from the data collection are added. This can entail qualitative data on observations, complaints from the farmers, weather conditions, harvest issues, road conditions, the sampling procedure and how the sample was allocated etc. These insights help to interpret quantitative data delivery in a better way.
 4. **Handover meeting:** Once the data is delivered, Akvo will set up a call with the SDM analyst responsible for the case, the analyst from Akvo responsible for data delivery, and the Akvo colleague that was responsible for the data collection and enumerator training. During this meeting, the data delivery report is discussed and the Akvo team can answer any questions the SDM analysts may have about the report, the data collection or the context.













5. Timeline and teams

Each primary data collection follows a sequence of steps to ensure timely data collection in a cycle of eight weeks from the moment of the kick-off meeting to the company visit of IDH. In practice, however, the timelines vary and on average take six to eight weeks. The table below shows the timeline and teams involved in the primary data collection process.

● IC team ● Akvo ● TA team ● SDM team ● Company

- Main team responsible for activity
- Other teams involved

Kick-off & Design	Week 1	● ● ● ● / ● ● ●	Clarify data needs Introduction call for new case
	Week 2	● / ● ● ● ● ● ● ● ● ●	Kick-off call Share intake form

	Week 3	 	Submit intake form Start survey design
	Week 4	  	Survey design Sample design and allocation Plan logistics data collection
Data collection	Week 5	  	Enumerator training Data collection Monitor data collection
Data cleaning	Week 6		Data cleaning
Reporting & dissemination	Week 7	 	Data delivery Data handover meeting
	Week 8		Country visit IDH

To ensure timely data collection, some of the success factors to consider include;

- Clear communication and understanding of what is needed and roles of IDH, the company and Akvo during the primary data collection
- Timely and clear response to requests by the SDM company e.g in sharing the farmer database and input form
- Enough time allocation for the PDC cycle from kick-off to data delivery
- Clear understanding on permit application requirements/needs and assistance with the application by the company and/or IDH

6. Ethical Considerations

The data collector abides by the principle of do no harm. Farmers are given an explanation on how the data will be used and are given the option to participate in the survey or not through an informed consent question that is built into the survey for each interviewee. (i.e. farmer or female decision maker in the household).

Furthermore, the data in the survey tool is only accessible to data collector team members that either built the survey or analyse the data. The analysed data is shared with IDH in two formats: raw, only accessible to the MEL and data specialist from Farmfit, and in an anonymised and cleaned version

that is accessible to the Farmfit team. The raw data is not shared with the SDM company as per the data sharing agreement with IDH and Akvo.

Data is stored in its original format on a cloud based database as a service from the data collector. After cleaning of the data the data can be accessed on the IDH Farmfit data portal. The personal data is made available for only two staff members of IDH. An anonymized data delivery file can be accessed by the SDM team. Access to this file is managed with different roles on the portal. Please refer to the [data anonymisation protocol](#) in annex 1 for more information on the anonymised data file.

The data collector has signed a non-disclosure agreement and a data processing agreement with IDH. The data collector is not allowed to share any of the data nor the farmer databases of the SDM company with any third party and does not have ownership of the data at any point in time. Enumerators sign a non-disclosure agreement with the data collector in which they state they will not share any data with third parties and will delete all information related to the farmers after the end of the assignment. If IDH chooses to stop working with the data collector, the data in the cloud based database will be deleted once all data has been transferred to IDH.

7. Guiding Principle for Updates to Methodology

While the intention is to maintain the Question Library as lean and standard as possible, questions may need to be adjusted or added at different times to ensure the data collected is fit-for-purpose. Questions that need to be added to a survey will be discussed between the company, IDH and Akvo (or other data collector) during the design phase, particularly when the intake form is completed.

Any additional questions will be added to the survey where they least interfere with the flow of the survey and will be labelled with the variable name "*new_[company reference name]_[theme/topic related to new question]*". To maintain version control, IDH and Akvo will review the question library once every quarter. The reliability of the questions will be analyzed by Akvo's data science team prior to the quarterly meeting so that the addition of the question to the standard Farmfit PDC question library can take the quality of the data produced by the question into account.

A question library gatekeeper will ensure changes are adequately documented and addressed in all relevant documentation.

ANNEXES

Annex 1: Data Anonymization Protocol

1. Anonymisation of data for Farmfit Primary Data Collections

Introduction

Within the Farmfit programme, the unit of analyses are the farmer and, when applicable, the female head of the household. Personal details are collected to, for instance 1: identify, by retrieving the geo location, whether data is collected in the agreed area and 2: to enable follow ups with the farmers by asking for a name and a phone number. However due to GDPR, in the current setup, these personal details should only be accessible by specific IDH employees and Akvo employees situated in European Union. Moreover, it is always favorable to only share personal details if there is a logical reason to do so. Therefore, this document clarifies how Akvo takes care of anonymising the data.

Anonymisation via R-code

For the data delivery, Akvo wrote an R-code (available in the private Git folder of IDH and Akvo) that outputs cleaned datasets including calculation of specific statistics. Moreover, the R-code leads to two datasets: the full, raw data and a dataset *excluding* the personal data. To delete the personal data, the steps below are taken:

Step 1 - Question groups

The questions groups holding personal data are purposely named 'Personal information & monitoring' (containing name and phone number) and 'Location - Confidential' (containing name of e.g. district, village, ward). The variables in these two groups are pulled and filtered from the dataset.

Step 2 - Variables on geolocation

All variables containing the word 'geolocation' are also filtered from the dataset.

Step 3 - Manual final checks

During and after running the entire code for data delivery, the designated analyst from Akvo always checks whether the columns holding personal data are indeed removed from the cleaned dataset that should *not* hold any personal data. The column headers are standardised variable names such as `spi_name_of_farmer` or `pi_mobile_number_farmer`. The variable names remain consistent amongst all data collections, independent of survey language. Therefore, the analyst can easily check, by reading the column headers, whether the columns containing personal data are deleted from the dataset.

```

# Select variables with private information
private_info <- survey_questions %>%
  filter(section %in% c("personal information & monitoring", "location - confidential")) %>%
  filter(variable != "geolocation") %>%
  pull(variable)

# Select Geolocation separate
geolocation <- names(Data)[names(Data) %like% "geoLocation"]

# Remove columns from set
Data <- dplyr::select(Data, select = -c(private_info, geolocation))

```

The R-code for Step 1 and 2

Note: Since the locations are used to calculate the PPI scores, all location based information is filtered from the dataset after the PPI scores have been retrieved.

Akvo Flow set up

As of June 2020, a new feature is added to the data collection tool Flow. All question types holding personal data can be marked. This way, when data is marked as personal data it can be hidden when exporting the data with the Flow API.

2. Joint Controller Data Sharing Protocol

Introduction

Akvo has been supporting IDH with the collection of primary farmer data. This data contains personal data of farmers. Because Akvo is an EU based company this data needs to be protected under GDPR regulation. Under the current contract IDH is the Data Controller and therefore responsible for the security and means of processing of the data, and Akvo is the data processor. Furthermore, IDH, as a company based in the EU who is commissioning the data collection, is a Data Controller under GDPR legislation and thus needs to comply with it regardless of which organisation collects the data.

What is a Data Controller?

The data controller determines the purposes for which and the means by which personal data is processed. So, if your company/organisation decides ‘why’ and ‘how’ the personal data should be processed it is the data controller. Employees processing personal data within your organisation do so to fulfil your tasks as data controller.

What is a Joint Controller?

Your company/organisation is a joint controller when together with one or more organisations it jointly determines ‘why’ and ‘how’ personal data should be processed. Joint controllers must enter into an arrangement setting out their respective responsibilities for complying with the GDPR rules. The main aspects of the arrangement must be communicated to the individuals whose data is being processed.

What is a Data Processor?

The data processor processes personal data only on behalf of the controller. The data processor is usually a third party external to the company. However, in the case of groups of undertakings, one undertaking may act as processor for another undertaking.

Source: [European Commission](#)

What are the roles and responsibilities of a joint controller?

The roles and responsibilities of a Joint Controller are the same as the Data Owner, which in this case is IDH. As a Data Controller IDH is responsible for making sure that all processing of the data is performed in accordance with GDPR regulation.

How should they request data from Akvo?

This should be part of the initial project request that reaches Akvo through Monday.com. On Monday.com the appropriate IDH Business Units fill out a request form for a project which contains a section on data delivery. In the request form it should be clear that there is a third party with a Joint Controller contract.

How should Akvo share data with joint controllers?

Secure file sharing - request Valeria

Other obligations of Akvo towards joint controllers

Akvo has no other obligations to joint controllers outside that agreed in the relevant Data Processing Agreement. Joint controller does not exclude Akvo from being contracted by the Joint Controller to perform additional tasks related to the data set in question e.g. analysis, further cleaning etc.

Next steps

- A. Request additional field for joint controllers in request form Monday.com.
- B. Choose and agree upon a file sharing system.

Annex 2: Enumerator Field Guide

Informed consent

Thank you for the opportunity for us to speak with you. We are a research team from Akvo and we are working together with **SDM company**. We are collecting data on behalf of IDH. IDH is an international organization that seeks to improve the services that farmers receive. We are conducting a survey to learn about farmer services, practices and income in both **focus crop** production and wellbeing of households in **location**. The information we collect will help IDH to review and plan activities with both focus crop farmers and **SDM company** which offers different services to you and other **focus crop** farmers in this area. You have been selected to participate in the interview based on your relationship with **SDM company**. I would like to ask you some questions about your household and your farm. The questions usually take about 30-45 minutes. All of the answers you give will be anonymous (your name will not be tied to the answers) and will not be shared with anyone other than members of our survey team and IDH. IDH might use the data for continued research for improving smallholder farmer agriculture around the globe. Your participation is voluntary, but we hope you will agree to answer the questions since your views are important and could help improve the services farmers receive. If I ask you any questions you don't want to answer, just let me know and I will go on to the next question or you can stop the interview at any time.

If you have any questions in future regarding the study and/or the interview, or concerns or complaints you can contact **SDM manager** mobile number phone number from IDH

Correspondent

Who to interview? Interview the person in charge of the farm. This would be the person who owns the land or is the caretaker (in the absence of the actual owner).

Farmer

- **A small-holder farmer:** *2 hectares*³ of land and below
- If a farmer farms on more than *2 hectares*, please **stop** the interview and continue to the next farmer.

Farm

The size of the farm is the size the farmer has grown his/her crops only.

Multiple farm plots: When asking about farmsize ask the small-holder farmer if he has multiple plots and add up all plots to get to total farm size. If it's bigger than *2 hectares* of land, the farmer is not considered a small-holder farmer.

Measurement units:

Plot size: e.g *Hectares or acres*

Weight: *Kilograms (kg)*

³ The maximum number of acres or hectares for a farmer to be considered as small holder is determined on a case by case basis by asking the company what they consider a smallholder.

Most common packaging unit for rice seeds and rice grains: *Bags of 100 KG or 300 KG*

Contextual information:

Farmgate price per *Packaging type (e.g. 30 kg bags)*: min. *100 currency X*, max. *200 currency X*

Farmgate price per *Packaging type (e.g. 50 kg bags)*: min. *100 currency X*, max. *200 currency X*

Number of harvest seasons in a year: *One* harvest season

Yield per acre of crop: min. *1500 kg/hectare* max. *3500 kg/hectare*

Reasons for crop loss: *Flooding, non-certified seeds*

Note:

All questions about labour, equipment and inputs refer to the SDM crop. We don't need to know labour, equipment and inputs for other crops the farmer grows.

FARM

Size of the farm: Be sure to double-check in which unit of measurement the farmer is answering the question (acres, hectares, etc.). In *Country Z*, farms are usually measured in *hectares*.

The answers can be: 0.5 - 2 hectares for small-holder farmers

***The tool tip of this question:** Make sure not to include the house of the farmer. Enter the number '9999' if the answer is interpreted as 'I don't know' and '9998' if 'I prefer not to answer.'*

CROP REVENUE

Calculation of harvest seasons: Crops can have multiple harvest seasons per year. If the farmer has 2 harvest seasons, please make sure to repeat this question group.

Size of farm dedicated to crop: Use the same unit of measurement as the farmer used to answer the total size of the farm. In case the farmer does not know the size:

- 1) Ask to give an estimation by probing: was it a quarter of the total size of the land, half of the land, etc.?
- 2) Check the question on total size of the land (second question under 'Farm') and do the maths.

Example: Farmer says half of the land was dedicated to rice. Under question group ‘Farm’ it was answered that the total size of the farm is 2 acres. Thus, you fill in that 1 acre of land was dedicated to rice. **Note:** Farm size dedicated to SDM crop can’t exceed total farm size.

Lost crop: For each season, the farmer needs to indicate how much of their main crop has been lost. In case the farmer is not able to answer directly on the loss, probe by asking about drought, floods, birds, disease, and pests.

OTHER CROP INCOME | LIVESTOCK/POULTRY INCOME | OFF-FARM INCOME

Calculating additional income for last 12 months: The questions on non-SDM crops, livestock and off-farm income will be answered in a **time frame of 12 months** instead of (harvest) seasons. For this case, you need to ask additional income from *October 2018* to *September 2019*. In case the farmer does not know exactly, help the farmer to calculate per month/quarter/season and add it up to a total of 12 months. Fill in the numbers from the local currency value.

LABOUR COST

Number of people working on activities: For each activity, make sure you note the answer of number of people involved **excluding** the farmer him/herself.

Number of days and payment per person per activity: You will find that farmers in *Country X* often hire people per day. In that case, ask for the number of days one person works on an activity per day and multiply.

For all the "how many days questions" ensure clarity that we are asking how many days, an avg hired person spent working for you throughout the year for this purpose.

How much did you pay the people you hired per person per month?: Enumerators will input the amount in numbers of local currency

OFF-FARM INCOME

Sources of income that do not relate to crop or livestock: Next to income from farming activities, some farmers will have income generated through other means:

Example remittances: Family member, relative or friends sends money to complement the farmer’s income.

Example gift: Farmer receives a sum of money that he/she does not have to re-pay.

LOANS

Example in-kind loan: Farmer buys seeds on a loan from a seed company. The loan will be deducted when the seed companies buys products from the farmer.

Annex 3: Policy document: Monitoring with enumerators during COVID 19

Guiding principles:

- Ø We follow the principle of 'do-no-harm'. Specifically, staff + partners + enumerators + individuals well being is our #1 priority.
 - o Safety of Staff
 - o Safety of Partner
 - o Safety of Enumerators (also during travel to location)
 - o Safety of individuals and beneficiaries

- Ø We always follow country specific government (COVID-19) regulations:
 - o We have a tailored response that is country specific.
 - o When working from home is suggested for non-essential work, we interpret that to also mean non-essential field work.
 - o The direct support to monitoring the impact or response to Corona related monitoring efforts is deemed essential, and will be prioritized.

- Ø We focus on appropriateness and a critical needs assessment in cases where we are allowed and able to operate:
 - o If possible, we shift to remote monitoring or alternative models for data collection
 - o If not, we follow health and safety regulations
 - o We aim to minimize contact points with beneficiaries

- Ø We follow WHO health and safety guidelines:
 - Wash your hands frequently
 - Maintain social distancing (at least 1,5 meter)
 - Practice respiratory hygiene
 - If you have fever, cough and difficulty breathing, stay home and seek medical care early.
 - If you have been in contact with a person that has been positively diagnosed, you stay in quarantine for 14 days

- Ø We stay informed, take-in, follow and communicate local advice

- Ø Our TC team fast-tracks and suggests alternative models for data collection so our partners can continue to offer vital development services. We clearly communicate the trade-offs.

Ø We verify and improve this policy in cooperation with key strategic partners, and based on the latest advisories from respective countries and relevant multilateral agencies (WHO, UNICEF, etc..)

