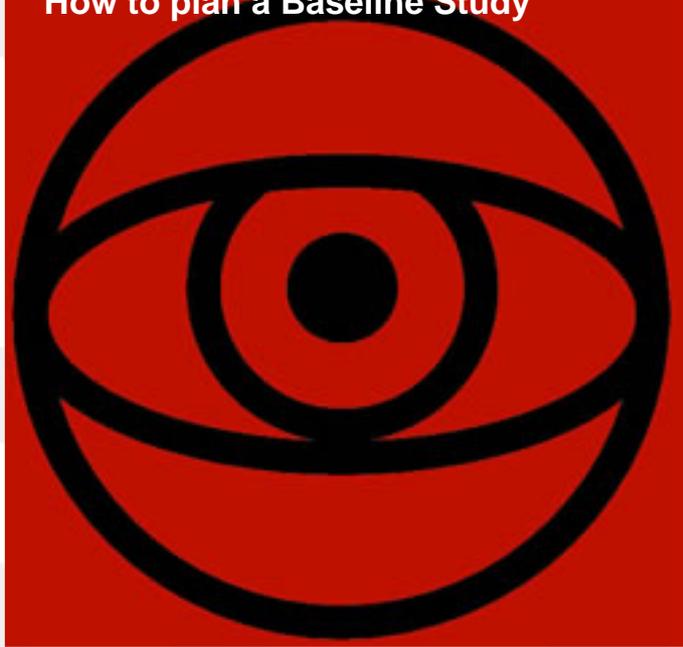


How to plan a Baseline Study



Monitoring & Evaluation Guidelines



United Nations World Food Programme
Office of Evaluation

- What is a Baseline Study 5
- When to do a Baseline Study 7
- How to check the Link between the Operation Design and the M&E elements 10
- Types of Data – Quantitative versus Qualitative 15
- What are the Sources and Uses of Primary and Secondary Data 17
- Selecting the Unit of Study 19
- Sampling 20
- What is meant by Disaggregating or Stratifying and how is it done 23
- Comparison Groups: why not to use Them 25
- Preparing the Baseline Work Plan and Budget 27

How to plan a Baseline Study

Overview

Introduction. The purpose of this module is to explain how to plan a baseline study.

Why is this Module important?

This module is important because it explains what a baseline study is and how it is related to the mid-term and terminal evaluations. It describes when to conduct a baseline study and how to ensure that the outcome and impact indicators on which the study will collect data are clearly stated and, not least, in line with the operation design as outlined in the logical framework.

The module defines key concepts related to baselines and provides guidance on the use of primary or secondary data, sampling techniques, how to define the unit(s) of study, whether to use comparison groups, etc. In addition, it describes what a WFP baseline study work plan and budget should look like.

What does this Module aim to achieve?

This module has the following objectives:

- To explain what a baseline study is and how, as part of a larger monitoring and evaluation (M&E) strategy, it forms the basis for measuring change over time to an operation's outcome and impact indicators;
- To explain the type of operations for which baseline studies are required and the appropriate timing of a baseline study in relation to an operation's programme cycle;
- To describe the critical relationship between M&E and operation design, and explain what steps should be followed to ensure that this link is clearly articulated;
- To describe the 2 general types of data – quantitative and qualitative – that can be used for M&E purposes;
- To describe 2 broad categories of data sources – primary and secondary – and the appropriate use of each in providing information for use in the monitoring and evaluation of WFP operations;
- To identify the appropriate unit(s) of study and explain why it is important to specify it (them) in indicators;
- To explain what sampling is, and describe when to use probability and non-probability sampling;
- To explain what stratifying and disaggregating mean in relation to sampling, data collection, indicators and analysis, including the requirements for monitoring and evaluating WFP's Commitments to Women. In addition, to provide the rationale for stratifying prior to data collection, as well as during analysis.
- To explain the rationale for using comparison groups, and why it is inappropriate for the majority of WFP operations to do so;
- To describe what should be included in a baseline study work plan and budget.

What should be reviewed before starting?

- What is RBM Oriented M&E
- How to design a Results-Oriented M&E Strategy for EMOPs and PRROs
- How to design a Results-Oriented M&E Strategy for Development Programmes

Section Titles and Content Headings

- **What is a Baseline Study**
 - Introduction
 - What is a Baseline Study
 - Before and after Evaluation Design
 - An Example of how a Baseline Study can be used to measure Change over Time
- **When to do a Baseline Study**
 - Introduction
 - Baseline Requirements in WFP
 - When to do a Baseline Study
 - An Example of underestimated Impact associated with delayed Baseline Data Collection
- **How to check the Link between the Operation Design and the M&E elements**
 - Introduction
 - The Standard Logical Framework Matrix and how it relates to M&E
 - The main Contents of the Logical Framework Matrix
 - How to check the Design Logic in a Logical Framework
 - How to check the M&E Elements in a Logical Framework
 - An Example of how to check the Design Logic in a Logical Framework
 - An Example of Distinct and Separate Results Hierarchy Levels and Design Elements
 - An Example of SMART Indicators within Each Level of the Results Hierarchy
- **Types of Data – Quantitative versus Qualitative**
 - Introduction
 - What are the Characteristics of Quantitative and Qualitative Data
 - Examples of Quantitative and Qualitative Data
- **What are the Sources and Uses of Primary and Secondary Data**
 - Introduction
 - What are the Differences between Primary and Secondary Data
 - Appropriate Uses of Primary and Secondary Data
 - An Example of using Secondary Data in Development
 - An Example of a Secondary Data Source for Emergency Operations (EMOPs)
 - An Error to avoid
- **Selecting the Unit of Study**
 - Introduction
 - What is a Unit of Study
 - Examples of Units of Study
 - Examples of Units of Study in a Nutrition Programme
- **Sampling**
 - Introduction
 - What is Sampling
 - What distinguishes Probability Sampling from Non-probability Sampling
 - Examples of Non-probability and Probability Sampling for a Baseline Survey
 - An Example of an Estimate from a Probability Sample
- **What is meant by Disaggregating or Stratifying and how is It done**
 - Introduction
 - What is Stratification and what is Disaggregation
 - Stratification Requirements for the M&E of WFP's Commitments to Women

- Example of Pre-stratification of a Probability Sample
- Example of Stratification of a Non-probability Sample
- Examples of the Stratification Factors that are listed in Indicators
- **Comparison Groups: why not to use Them**
 - Introduction
 - What is a Comparison Group
 - Comparison Groups: why not to use Them
 - An Example of the Difference between using and not using Comparison Groups
- **Preparing the Baseline Work Plan and Budget**
 - Introduction
 - Proposed Content for Baseline Study Work Plan and Budget

What is a Baseline Study

Introduction. This section explains what a baseline study is and how, as part of a larger M&E strategy, it forms the basis for measuring change over time in an operation's outcome and impact indicators. The concept of measuring change over time at the outcome and impact level forms the backbone of a results-based management (RBM) approach to programming.

What is a Baseline Study

A baseline study simply defines the 'pre-operation exposure' condition for the set of indicators that will be used to assess achievement of the outcomes and impact expressed in the programme's logical framework. When compared with the condition of the same indicators at some point during implementation (mid-term evaluation) and post-operation implementation (final evaluation), the baseline study forms the basis for a 'before and after' assessment or a 'change over time' assessment. Without baseline data to establish pre-operation conditions for outcome and impact indicators it is difficult to establish whether change at the outcome level has in fact occurred.

Before and after Evaluation Design

In the design of a before and after evaluation, baseline studies are a critical element in the formula for measuring change over time.

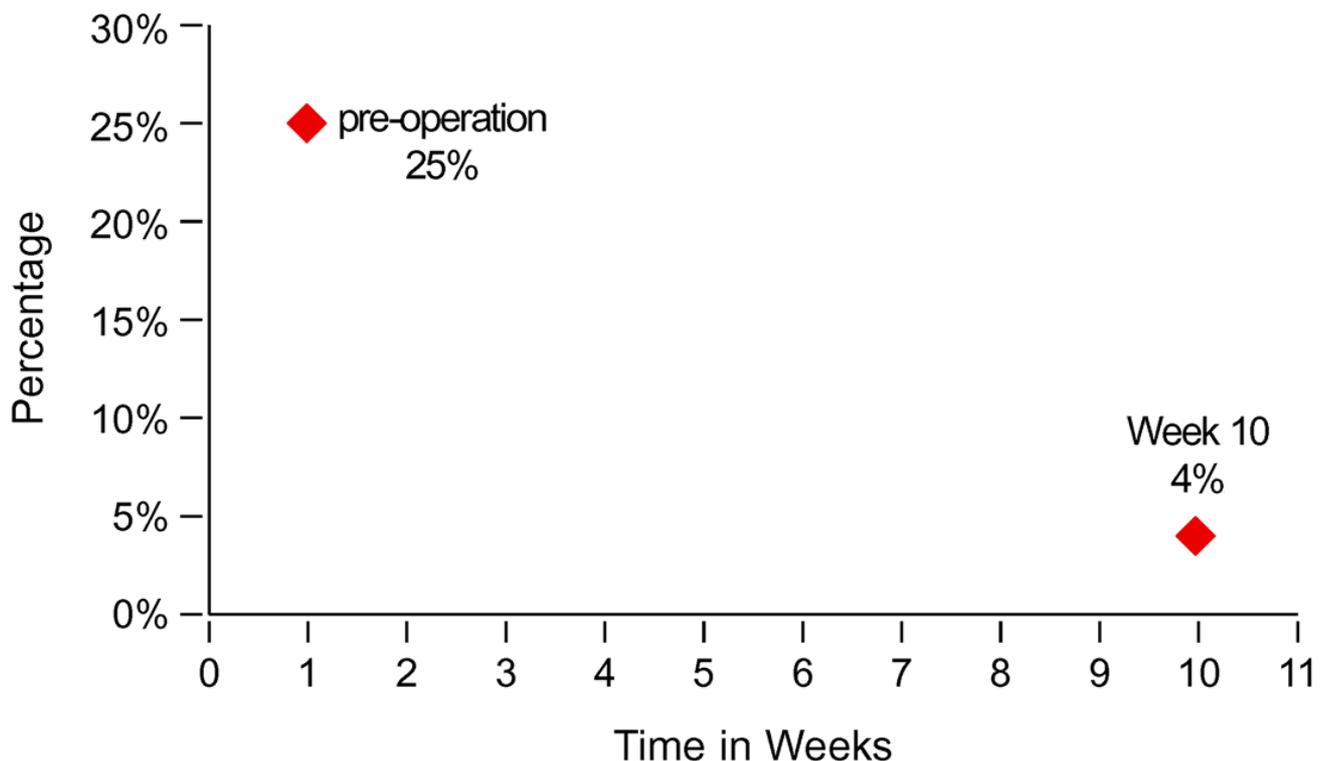
Pre-programme condition for outcome and impact indicators	=	Post (or mid-term)-programme condition for outcome and impact indicators	-	Change over time in outcome and impact indicators associated with WFP activities
-----------------------------------------------------------	---	--------------------------------------------------------------------------	---	----------------------------------------------------------------------------------

An Example of how a Baseline Study can be used to measure Change over Time

The example below shows how baseline data are used to establish pre-exposure conditions and estimate the change over time related to WFP operations. Although the example is for an EMOP, the same concept applies to development operations and PRROs (although the time line and indicator of interest will differ).

A pre-operation baseline study for an emergency feeding operation establishes an acute malnutrition (weight-for-height < -2 standard deviation) prevalence of 25 percent among children under 5 years of age, the key impact performance indicator for the operation. A subsequent study is conducted at the end of 10 weeks to assess the impact of the operation on the impact indicator, and yields a prevalence of 4 percent. The change in the key operation impact indicator is estimated to be approximately 21 percent (e.g. 25 percent from the baseline study minus the 4 percent found at the end of week 10).

Prevalence of Acute Nutritional Status (Weight-for-Height)



When to do a Baseline Study

Introduction. This section explains the type of operations for which baseline studies are required and the appropriate timing of a baseline study in relation to an operation's programme cycle.

Baseline Requirements in WFP

In WFP, a baseline study is required for every type of operation. However, the rigour of the methods used to establish baseline conditions varies according to the type of operation being implemented. A compromise must be reached between the need for robust, precise data to establish pre-operation exposure conditions and the cost of collecting such data in terms of resources (financial, human and time). Country Programmes that are focused on development should invest more resources and, as a result, conduct more rigorous baseline studies. PRROs and EMOPs will establish the necessary rigour for a baseline study by considering the available resources and the information needs. However, a minimum standard for establishing pre-operation conditions through baseline studies should be applied in all WFP operations.

The baseline study is just 1 component of the M&E design that outlines the planned M&E data collection and analysis. The entire evaluation strategy, including the design and budgeting of the baseline and subsequent studies (mid-term and final evaluations), must be developed during the planning or design stage of an operation.

When to do a Baseline Study

In relation to the programme cycle, a baseline study should be conducted prior to the onset of operation activities in order to establish the pre-operation exposure conditions of the outcome and impact level indicators. However, it is not uncommon for baseline studies to be conducted after activities have already begun.

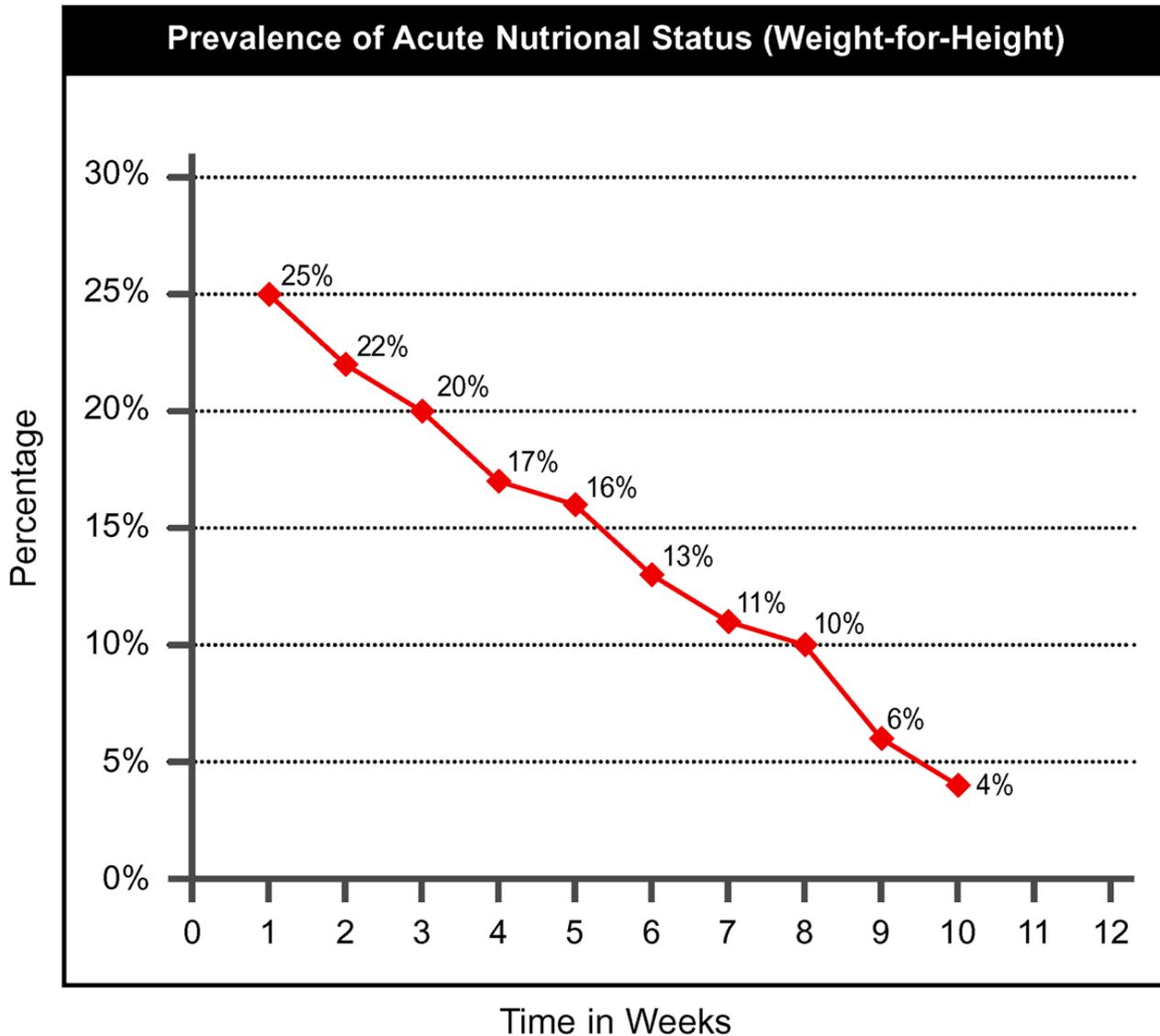
It should be noted that, for most operations, there is a delay between WFP's output delivery activities and their measurable effect on outcome and impact performance indicators. As a result, baseline studies will still provide an accurate estimate of pre-operation conditions even after the operation has begun, as long as the outcome and impact performance indicators have not yet been affected. However, this time lag varies from a few days to a few months, according to the type of operation and the environment in which it is being implemented. For many operations it is difficult to estimate exactly how long this time lag will be.

Delays in conducting baseline studies, especially when an operation's activities have already influenced the outcome and impact performance indicators, are costly and likely to lead to an underestimation of the operation's overall impact. WFP operations should therefore aim at conducting baseline studies before operation activities begin. When this is not possible, baseline studies must take a high priority and data should be collected very close to the beginning of the operation, at the latest.

In some cases when a baseline study has not been conducted, evaluators find themselves attempting to establish the change over time at the mid-term and final evaluations without the benefit of knowing the pre-operation conditions of the key indicators of interest. Retroactively constructed baseline conditions (a much weaker evaluation design) should only be used in situations where baseline data have not been collected and no other choice is available.

An Example of underestimated Impact associated with delayed Baseline Data Collection

The example below illustrates the potential underestimation of impact associated with delayed baseline data collection for an EMOP in which acute malnutrition was the impact performance indicator. The same concept can be applied to development operations or PRROs (although the time line between the baseline and subsequent evaluation will differ).



Week in which baseline data was collected	Baseline estimate	Subsequent estimate at week 10	Estimate of change over time (baseline estimate minus week 10 estimate)
1	25%	4%	21%
2	22%	4%	18%
3	20%	4%	16%
4	17%	4%	13%
5	16%	4%	12%

Week in which baseline data was collected	Baseline estimate	Subsequent estimate at week 10	Estimate of change over time (baseline estimate minus week 10 estimate)
6	13%	4%	9%
7	11%	4%	7%
8	10%	4%	6%
9	6%	4%	2%
10	4%	4%	0%

How to check the Link between the Operation Design and the M&E elements

Introduction. This section clarifies the critical relationship between M&E and operation design, and provides the steps to follow to ensure that this link is clearly articulated.

The Standard Logical Framework Matrix and how it relates to M&E

The primary purpose of M&E is to measure the degree to which an operation design is implemented as planned and how successfully it achieves its intended results. The operation design describes how inputs and activities will result in outputs delivered by WFP and its partners, and how the operation designers believe these outputs will, in turn, result in desired outcomes and impacts.

The relationship between each of these levels is described in a logical framework hierarchy for the operation and represents a hypothesis concerning how the operation, starting with the initial resources or inputs that are available, will bring about the desired results. When a results-based approach to design is used, the desired outcomes or impacts are identified first, then the outputs needed to achieve those outcomes, and then the inputs and activities needed to deliver those outputs.

The logical framework approach produces a matrix (see following page), which combines the concepts of results-based management (RBM), results-based operation design and M&E.

What the operation will do; what it seeks to achieve	How performance will be measured		Factors outside management control and that may affect project performance
Logical framework hierarchy	Performance indicators	Means of verification	Assumptions and risks
Impact	(Impact)		
The higher objective to which this operation, along with others, is intended to contribute	Indicators (increasingly standardised) to measure programme performance	The programme evaluation system	Risks regarding strategic impact
Outcome	(Outcomes)		
The outcome of an operation; the changes in beneficiary behaviour, systems or institutional performance caused by the combined output strategy and key assumptions	Measures that describe the accomplishment of the outcome; the value, benefit and return on the investment	People, events, processes and sources of data for organising the operation's evaluation system	Risks regarding programme-level impact
Outputs			
The actual deliverables; what the operation can be held accountable for producing	Output indicators that measure the goods and services finally delivered by the operation	People, events, processes, sources of data – supervision and monitoring system for validating operation design	Risks regarding design effectiveness
Activities	Inputs/resources		
The main activity clusters that must be undertaken in order to accomplish the outputs	Budget by activity; monetary, physical and human resources required to produce the outputs	People, events, processes, sources of data – monitoring system for validating implementation progress	Risks regarding implementation and efficiency

The main Contents of the Logical Framework Matrix

Each of the 4 columns in the Logical Framework is described in the following paragraphs. The first and fourth columns articulate operation design and assumptions, while the second and third columns outline the M&E performance measurement indicators and means in order to test whether or not the hypothesis articulated in the operation design holds true.

Column 1: This column outlines the design or internal logic of the operation. It incorporates a hierarchy of what the operation will do (inputs, activities and outputs) and what it will seek to achieve (purpose and goal).

Column 2: This column outlines how the design will be monitored and evaluated by providing the indicators used to measure whether or not various elements of the operation design have occurred as planned.

Column 3: This column specifies the source(s) of information or the means of verification for assessing the indicators.

Column 4: This column outlines the external assumptions and risks related to each level of the internal design logic that is necessary for the next level up to occur.

How to check the Design Logic in a Logical Framework

To check the design logic of the logical framework, review and test the internal and external logic (columns 1 and 4, respectively) and the feasibility of the operation's logical framework. Test the logic beginning with inputs and move upwards towards the impact using an "if" (internal logic) "and" (external logic) "then" (internal logic at the next level) logic test. Where necessary, adjust the logical framework to overcome logic flaws or unfeasible/unlikely relationships among various levels of the logical framework hierarchy. If no logical framework exists for the operation, consult the Logical Framework Guidelines.

Specifically check that the following conditions hold:

- Inputs are necessary and sufficient for activities to take place
- Activities are necessary and sufficient for outputs that are of the quality and quantity specified and that will be delivered on time.
- All outputs are necessary, and all outputs plus assumptions at the output level are necessary and sufficient to achieve the outcome.
- The outcome plus assumptions at the outcome level are necessary and sufficient to achieve the impact.
- The impact, outcome, and output statements are not simply restatements, summaries or aggregations of each other, but rather reflect the resulting joint outcome of 1 level plus the assumptions at that same level.
- Each results hierarchy level represents a distinct and separate level, and each logical framework element within a results hierarchy level represents a distinct and separate element.
- The impact, outcome, activities, inputs and assumptions are clearly stated, unambiguous and measurable. Impacts and outcomes are stated positively as the results that WFP wishes to see. Outputs are stated positively in terms of service/product delivery.
- The assumptions are stated positively as assumptions, rather than risks, and they have a very high probability of coming true.

How to check the M&E Elements in a Logical Framework

Check that the following conditions hold in the logical framework:

- Indicators for measuring inputs, activities, outputs, outcome and impact are specific, measurable, accurate, realistic and timely (SMART) (column 2).

- Beneficiary contact monitoring (BCM) indicators are identified for the purpose of tracking progress between outputs and outcomes and are noted at the outcome level.
- 2 levels within 1 logical framework do not share the same indicator (if they do, the indicator at 1 level is not specific enough to that level or the design logic between levels is flawed).
- The unit of study (e.g. individuals, children, households, organisations) in the numerator and, where applicable, the denominator of each indicator are clearly defined such that there is no ambiguity in calculating the indicator.
- The means of verification for each indicator (column 3) are sufficiently documented, stating the source of the data needed to assess the indicator (be sure that sources of secondary data are in a useable form).

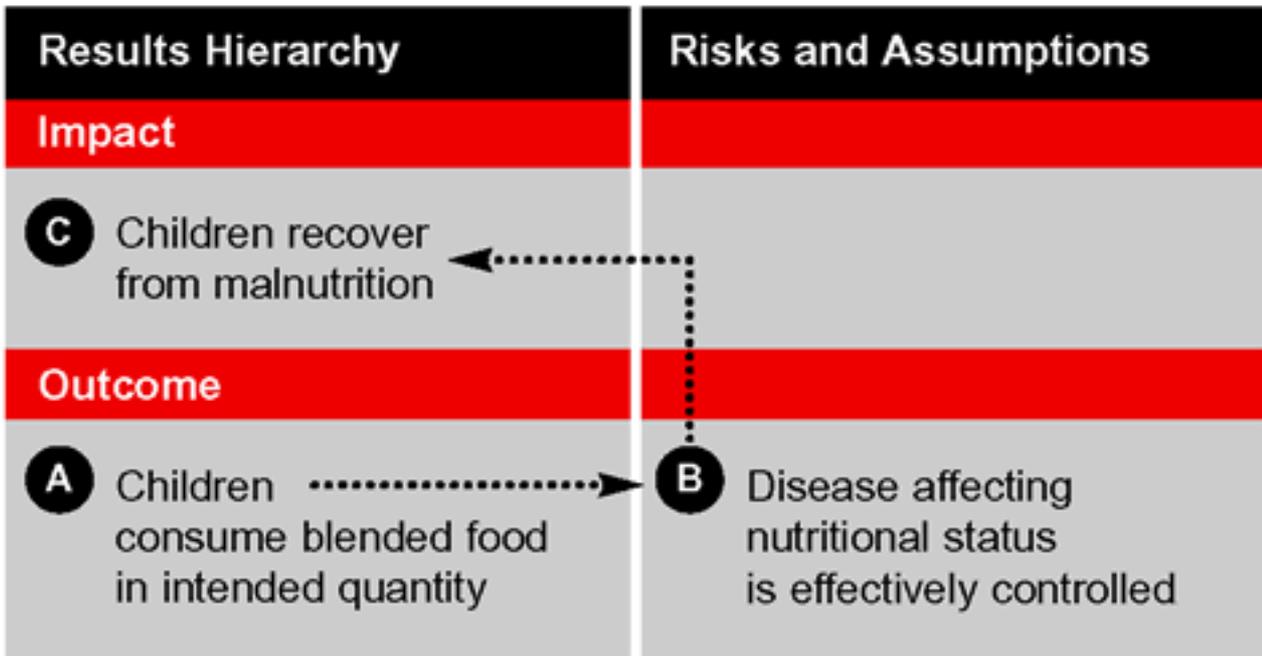
An Example of how to check the Design Logic in a Logical Framework

The following diagram is an example of testing the internal and external logic of a nutrition project’s logical framework using the if-and-then logic test.

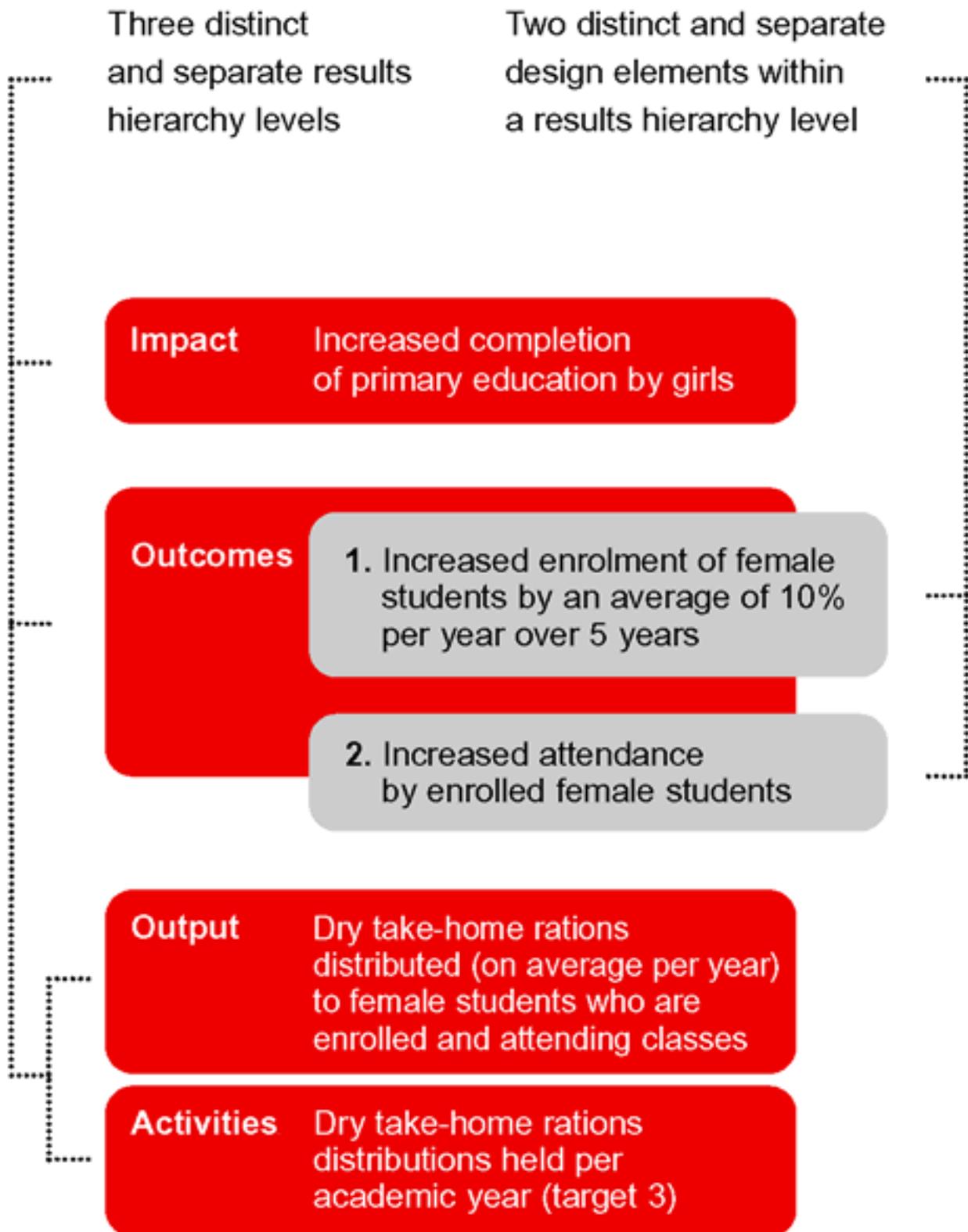
If we deliver the outputs through our planned activities and using the planned inputs, and our assumptions at the output, outcome and impact levels hold true, **then** the desired outcome will occur and lead to the desired impact.

A + **B** = **C**

A without **B** does not equal **C**



An Example of Distinct and Separate Results Hierarchy Levels and Design Elements



An Example of SMART Indicators within Each Level of the Results Hierarchy

Once the design levels have been clarified and simplified, choosing SMART indicators is a relatively straightforward task. Note how each indicator is designed to measure the element in the corresponding row only.

	Results hierarchy	Performance indicators
Impact	Increased completion of primary education by girls	A. % of students dropping out between one grade and another by gender and grade (between grades) B. Numbers of girls completing grade 6 and grade 9
Outcome	Increased enrolment of female students by an average of 10% per year over 5 years	A. Number of girls enrolled at the beginning of each academic year
	Increased attendance by enrolled female students	A. % of girls absent for 3+ days/month
Output	Dry take-home rations distributed (on average per year) to female students who are enrolled and attending classes	A. Number of rations distributed to girl students per semester/per academic year

Types of Data – Quantitative versus Qualitative

Introduction. This section describes the 2 general types of data – quantitative and qualitative – that can be used for M&E purposes.

What are the Characteristics of Quantitative and Qualitative Data

2 general types of data exist – quantitative and qualitative – although the distinction between the 2 is often blurred. While quantitative data have long been cited as being more objective, and qualitative data as more subjective, more recent debates have concluded that both types of data have subjective and objective characteristics. As qualitative and quantitative data complement each other, both should be used.

Characteristics of Quantitative Data

Characteristics of Quantitative Data:

- Seek to quantify the experiences or conditions among beneficiaries in numeric terms.
- Use closed-ended questions with limited potential responses.
- Normally ask women, men, boys and girls to respond to questions on the basis of their individual experiences, or the experiences of their households.
- Often, but not exclusively, employ probability sampling techniques that allow for statistical inference (or estimation) to a larger population with defined levels of probability (or confidence) and tolerable error (or confidence interval); although not as complicated as often thought, determining the appropriate parameters for calculating sample size is likely to require some expertise.
- Use measurement techniques (e.g. measuring land area; maize yield, by weighing bags of maize; food consumption, through weighing food quantities to be consumed by type; anthropometric indicators of children).

Characteristics of Qualitative Data

Qualitative data seek to uncover the context, perceptions and quality of, as well as opinions about, a particular experience or condition as its beneficiaries view it. Data collection methods are more likely to employ a more participatory approach through the use of open-ended questions that allow respondents to expand on their initial answers and lead the discussion towards issues that they find important. These more participatory methods will commonly be used in the M&E of WFP operations. Sampling techniques for these methods are often purposive. Even when samples are selected randomly, these methods rarely require the rigorous determination of sample size, and respondents are often asked to generalise about the condition or experience in the larger population, rather than talk about themselves.

Examples of Quantitative and Qualitative Data

Quantitative	Qualitative
The mean amounts of food commodities remaining in sampled houses 1 week after distribution was 45 kg of maize and 2 kg of vegetable oil.	Most households have used up the majority of their monthly ration in the first week after delivery because they are expected to share the ration with neighbours who are not eligible.
38% of households have an income of less than 300 Kenyan shillings per month.	According to women in the focus group discussion, the majority of households do not have enough income to meet all of their food purchasing needs.
40% of children under 5 years of age are wasted (< -2 standard deviation weight-for-height), 90% of wasted children have had diarrhoea in the last 2 weeks.	Women suggest that every child is malnourished at some time during the year and they attribute this to chronic diarrhoea.
The mean amount of time women take to reach the primary dry-season water source in Garissa district is 2.3 hours.	Women spend most of the daylight hours collecting wood, water and fodder for animals. They view this as the main obstacle preventing them from participating in other economic endeavours.
8 out of 10 women in the focus group discussion have more than 1 child under 5 years of age.	In the village, all the women between 20 and 45 years of age have at least 1 child under 5, and most have 2. The time spent in child care is the second largest obstacle to women's participation in economic endeavours.
58% of new arrivals indicated travelling 3 or more days to reach the refugee camp.	New arrivals in the refugee camp arrived exhausted having travelled for long distances, which they suggested resulted in many deaths along the way.

What are the Sources and Uses of Primary and Secondary Data

Introduction. This section describes 2 broad categories of data – primary and secondary – and the appropriate use of each in providing information for use in the M&E of WFP operations.

What are the Differences between Primary and Secondary Data

Data sources are listed in the third column of the logical framework matrix under the heading “means of verification”. While the indicator articulates what information will be collected, the means of verification identifies where that information will come from.

Primary Data

Primary data is data that is collected through the use of surveys, meetings, focus group discussions, interviews or other methods that involve direct contact with the respondents – women, men, boys and girls.

Secondary Data

By contrast, secondary data is existing data that has been, or will be, collected by WFP or others for another purpose. Secondary data may include WFP Vulnerability Analysis and Mapping (VAM) data, data from the mid-term or final evaluation of a previous phase of WFP operations, data collected by other organisations or the government of the country concerned, or data gathered by research organisations. Routine data collected by institutions participating in an activity (e.g. schools, health centres) are exceptionally good sources of secondary data which could not be replicated by primary data collection without prohibitive expense.

Distinction between Primary and Secondary Data

The critical distinction between the 2 types of data is that primary data is collected by WFP or someone who WFP has hired specifically for the purpose for which the data are required. Secondary data have been, or will be, collected for another primary purpose (e.g. all secondary data were or are primary data for another study), but may be used for “secondary” purposes related to M&E in WFP operations. Note that both primary and secondary data sources can yield quantitative or qualitative data.

Appropriate Uses of Primary and Secondary Data

The collection of M&E data, both primary and secondary, must focus almost exclusively on the indicators and assumptions identified at each level in the logical framework for the operation.

Secondary Data

The use of secondary data represents tremendous cost and time savings to the country office, and every effort should be made to establish what secondary data exist and to assess whether or not they may be used for the M&E of WFP operations. Primary data is often collected unnecessarily and at great expense simply because monitors or evaluators had not been aware that the data were already available. It is critical to invest the initial time and resources to investigate what data exist, what data collection exercises are planned for the future, and how relevant the existing data are for the M&E of WFP operations.

Primary Data

However, primary data collection is sometimes warranted. Although a review of secondary data sources should precede any primary data collection, existing data do not always provide the appropriate indicators or the appropriate disaggregation of indicators needed to monitor and evaluate WFP operations effectively. Even secondary data that provides the appropriate indicators and disaggregation of indicators may not be useful if the data is out of date and the situation is likely to have changed since they were collected. This varies greatly according to the indicator for which the data is being collected and its volatility. For example, school enrolment data that is 1 year old may suffice for establishing baseline conditions prior to a school feeding programme, but acute nutritional data (wasting) that is only a month old may no longer represent an accurate estimate of current conditions for that indicator.

Importance of Documenting Data Collection Methods

Clear documentation of the methods to be used to collect primary and secondary data must be developed during the planning stage of an operation. As data is collected, any variations from the planned data collection methods must also be documented. This ensures that data is collected in the same way at different points in time and by different people. This is critical for ensuring that the data is comparable, and improves the accuracy of assessing the changes over time associated with a WFP operation.

An Example of using Secondary Data in Development

The most common practice is to use a combination of primary and secondary data to complement each other. School feeding programmes will draw extensively on school records to meet M&E data needs. Although teachers keep records of attendance and enrolment primarily for purposes other than reporting to WFP, this information fits well with the data needed by WFP in order to assess the outcomes and impacts of a school feeding operation, and is therefore an ideal secondary data source.

An Example of a Secondary Data Source for Emergency Operations (EMOPs)

During the early stages of an emergency, the data gathered by the emergency food needs assessment (EFNA) should satisfy most of the immediate criteria for baseline data. Efforts should focus on ensuring that the data is reliable and representative. This exemplifies how data collected for 1 purpose can be used to serve another in a cost-effective way. This is especially true in the case of using assessment data for M&E purposes during EMOPs and PRROs.

An Error to avoid

A common error when using secondary data sources or collecting primary data is to collect too many data. This results from data collectors' tendency to collect all the data that is related to their own topics of interest rather than focusing on the specific data that is required for M&E. This often leads to a reduced amount of time available for data analysis and, ultimately, dilutes the value of the information produced.

Selecting the Unit of Study

Introduction. The purpose of this section is to explain what a unit of study is and why it is important to specify it in indicators.

What is a Unit of Study

The primary unit of study refers to the unit of interest defined in the M&E indicators listed in the operation's logical framework as measures of whether or not design elements occur as planned. Indicators must specify the unit of study clearly in order to ensure that the same unit can be applied in baseline and follow-up studies (mid-term and final evaluations). This is important because it ensures comparability at different points in time when measuring indicators. It should also be noted that any M&E data collection (either the compilation of secondary data or the collection of primary data) involving information that assesses more than 1 indicator may involve more than 1 primary unit of study.

Examples of Units of Study

Commonly found primary units of study used in the M&E of WFP operations include (but are not limited to) households, individuals, children, land, organisations, institutions (e.g. schools, hospitals), government departments and physical assets such as classrooms.

Examples of Units of Study in a Nutrition Programme

A baseline survey for a nutrition improvement operation may seek to improve household consumption at the outcome level (indicator: percentages of households that consumed more, less or the same amount of food as they did in the previous month) and improve the nutritional status of children under 5 years of age at the impact level (indicator: percentage of children under 5 who are less than -2 standard deviation weight-for-height). In this baseline study there are 2 primary units of study, the household for the outcome level, and children under 5 for the impact level.

Sampling

Introduction. This section explains what sampling is and describes when to choose probability and non-probability sampling. Choosing the appropriate sampling methods is based on: i) the data collection method being used in primary data collection; and ii) the degree of statistical rigour needed for extrapolating the sample estimate to the larger study population.

What is Sampling

Sampling occurs when a subset of the population (or other unit) under study is selected from the larger group (the entire population under study). By studying the findings from that sample (denoted as “n”) it is hoped that valid conclusions can be drawn about the larger population (denoted as “N”) from which the sample was taken. Sampling is commonly employed in order to avoid the expense and time associated with total enumeration of the population, as is done during a census.

Sampling is used to select respondents from among the larger population. Sampling makes it possible to analyse the impact of a WFP operation. Whether 2 focus group discussions are held to analyse the impact of a WFP operation in a geographic region or 1,500 households, in the same region, are selected at random, visited and asked questions from a questionnaire, sampling is used.

What distinguishes Probability Sampling from Non-probability Sampling

Sampling methods can be divided into 2 broad categories: probability sampling and non-probability sampling. Within each of these a variety of subcategories exist and a number of ways of selecting the sample can be used. Both probability and non-probability sampling methods seek to gather data that provide a fair representation of the larger population, although the definition of “representative” varies between the 2 methods.

Probability sampling methods rely on statistical theory as a basis for extrapolating findings from the sample population (n) to the larger study population (N). By contrast, non-probability sampling does not utilise statistical theory to support inference from a sample population (n) to the study population (N), but rather relies on a more subjective determination of the degree to which a sample “represents” the larger study population. The choice of which method to follow depends on the intended use of the information and the importance placed on objective (probability sampling) versus subjective (non-probability sampling) determination of how representative the sample is.

Probability Sampling

Probability sampling allows for statistical inference. Statistical inference makes use of information from a sample to draw conclusions (inferences) about the population from which the sample was taken. The estimates are representative of a larger population, from which the sample population is taken, at a known and quantifiable level of confidence or probability. Estimates are given in ranges, called confidence intervals, although they are often expressed as a point estimate +/- a number of percentage points. Probability sampling is almost exclusively used with quantitative data collection methods.

The essence of probability sampling is that each unit of study (e.g. household, individual, child) in the study population for which the estimate is desired must have an approximately equal probability for selection and inclusion in the sample. In order to ensure that this critical criterion

is met, an exhaustive sampling frame must exist or be created for the unit under study (households, individuals, children). A sampling frame is a complete list of all the potential units of study (e.g. households, individuals or children) in the population from which the sample will be taken.

In many countries, it is impossible to find an existing sampling frame at the unit of study level and it is too costly to construct one. In these cases, cluster sampling is used. Cluster sampling aggregates the unit of study into groups or clusters for which a complete or nearly complete list is available. Although cluster sampling is very commonly used, it is rarely employed appropriately. Expert guidance should be sought in applying cluster sampling and determining the appropriate number and proportional weighting of clusters.

Determining the appropriate sample size is based on a set of parameters concerning the degree of confidence desired in the estimate, the design effect of the sample, the degree of tolerable error and the proportion or mean estimates for the variable of interest. The sample size calculation includes an additional parameter when the desire is to measure change over time. Expert guidance should be sought in determining the appropriate sample size needed if probability sampling is being used.

Non-probability Sampling

Non-probability sampling also seeks to draw conclusions about the larger population under study through using a selected sample or subset of that population. However, in non-probability sampling the basis for doing so is not supported by the statistical theory of inference, as it is in probability sampling. Non-probability sampling is almost always used for qualitative data collection methods and can be used for quantitative methods for which statistical inference is not desired.

Because there is no effort to draw statistical inferences from a non-probability sample (either because the conclusions apply to the sample population only or because the inference to a larger population is not supported by statistical theory), there is no sample size calculation formula, as there is in probability sampling. It is also common to select and consult groups (case studies, focus groups) that are made up of a number of units of study; for example, when 5 focus groups of 10 respondents each are consulted, the total sample is 50 units of study or respondents.

Despite this more free-flowing approach to sampling, the desire to draw conclusions about the larger population does influence the sample size and the way in which the sample is chosen. The intent is to get a sample that is fairly representative of the geographic area and other important differentiating characteristics of the population under study (e.g. wealth groups, sex, age, livelihood). Choosing characteristics on which to stratify the sample requires thinking through which factors influence the variable(s) of interest in the study and ensuring that each important subgroup of the larger population is included in the sample population.

While the guidelines for sample size are less strict for non-probability sampling, a balance must be struck between the ideal number of interviews or discussions to hold and the resources available for doing so. This is particularly important in the use of time-intensive qualitative studies in which a single discussion may take several hours to conduct. The most common types of non-probability sampling methods used for M&E in operations are:

- Purposive sampling (choosing respondents based on the fact that they are likely to give the best picture of the phenomena you wish to inquire about).
- Random sampling (using a random method to select respondents).
- Opportunistic sampling (simply choosing respondents based on their availability to participate at the moment you arrive to collect data).

In general, purposive and random sampling will yield better data than opportunistic sampling.

Because of the in-depth nature of qualitative methods, the sample size (sites or study units) will necessarily be limited. However, the ability to draw conclusions about the larger population from

the sample population is enhanced as the sample size increases. Seek guidance from experts concerning sample size for the particular data collection method being used. The aim should be to maximize the sample size within the constraints of available resources and while maintaining the highest level of data quality possible.

Examples of Non-probability and Probability Sampling for a Baseline Survey

The aim of the baseline survey is to determine the average number of weeks of food shortage suffered by households in the region during the dry season. The following examples illustrate the application of a probability sample and a non-probability sample during a baseline survey to establish pre-operation conditions for this indicator.

Non-probability Sampling

A non-probability sample of respondents is chosen to participate in a focus group discussion concerning food shortage during the dry season. 5 villages are randomly chosen for inclusion and, within each of those villages, 10 women and 10 men are chosen to participate in gender-separate discussion groups.

Probability Sampling

It is determined that 210 households will need to be chosen at random from among all the households in the target area to participate in a household survey concerning food shortage during the dry season. Because no list of all the potential households is available, a 2-stage cluster sampling design is used.

In the first stage, 30 villages (clusters) are chosen from the 239 potential sample villages, which are weighted in proportion to their estimated size (big villages are weighted more than small villages so that all households have an approximately equal chance of being included in the survey).

Within each of the 30 villages selected from among all the potential villages (clusters), the United Nations Children's Fund (UNICEF) pencil spin method is used to select 7 households for inclusion in the survey. A pencil is spun at the village's mid-point and every other household is interviewed in a line in the direction in which the pencil is pointing until 7 households have been selected. If the end of the village is reached before 7 households have been selected, the pencil is spun again and a new direction is chosen. Again, every other household is selected for inclusion in the survey.

An Example of an Estimate from a Probability Sample

A probability sample of 210 mothers with children under 5 years of age is taken, and each mother included in the sample is asked whether or not each of her children under 5 has had diarrhoea in the last 2 weeks. Because some mothers have more than 1 child under 5, the total number of children referenced in the sample is 332. Of these, 154 have had diarrhoea in the last 2 weeks. Therefore, the diarrhoea prevalence point estimate for the sample population is 46 percent and the confidence interval surrounding the estimate is 40 to 52 percent, meaning that, at 95 percent confidence (the confidence level used to determine the sample size), the true population prevalence lies between 40 and 52 percent (e.g. 95 out of 100 samples in this range will contain the true population prevalence for diarrhoea).

What is meant by Disaggregating or Stratifying and how is It done

Introduction. This section explains what stratifying and disaggregating mean in relation to sampling, data collection, indicators and analysis, including the requirements for monitoring and evaluating WFP's Commitments to Women. In addition, it provides the rationale for stratifying prior to data collection, as well as during analysis.

What is Stratification and what is Disaggregation

Whenever a comparison is made between 2 groups – regardless of how the groups have been defined – the data regarding those groups are being stratified or disaggregated. In WFP, important factors for stratification include age group and gender, reflecting WFP's Commitments to Women. Other commonly found variables for stratification are geographic location (village, district, province, etc.) and wealth group. The factors by which data can be stratified (i.e. by splitting a larger group into 2 strata – e.g. men and women – or many strata – e.g. 5 age groups) are endless and should be selected on the basis of the analytic needs of the operation.

When stratification is carried out during the analysis stage it is known as disaggregation. The concept of stratification must be applied prior to data collection, while the sampling strategy is being devised, in order to ensure that enough data is collected about each of the groups that are being compared. If this does not occur, there may be too few sample units from 1 or more of the stratified groups to allow valid conclusions to be drawn.

The stratification concept is applicable to both probability and non-probability sampling. In probability sampling, a separate sampling frame is developed for each stratum (or grouping of units) and the selected sample size is applied to each stratum. The same rules apply in cases of non-probability sampling. Each time an additional stratum is added, the sample size doubles. The decision to carry out pre-stratification during sampling should therefore only be made after careful consideration of the additional costs associated with adding strata.

The best practice is to list the factors for stratification in the indicators. This ensures that critical pre-stratification needs are considered prior to choosing a sample. It also ensures that post-stratification (or disaggregating) occurs during analysis.

Stratification Requirements for the M&E of WFP's Commitments to Women

As part of its Commitments to Women, WFP has made a commitment to generate and disseminate data that is disaggregated by gender.

All WFP monitoring and reporting will specify:

- Men's and women's percentage shares of the resources received from food distribution (e.g. disaggregated output indicators);
- Men's and women's shares of benefits by category of activities (e.g. disaggregated beneficiary contact monitoring [BCM], outcome and impact indicators);
- The percentage of positions held by women in the planning and management of food distribution (e.g. disaggregated activity and output indicators).

All M&E information must assess whether WFP's gender commitments have been adhered to and the reasons for any failure to do so. All indicators must also be computed to reflect WFP's gender commitments (i.e. they should be disaggregated by gender).

Example of Pre-stratification of a Probability Sample

Separate estimates of acute malnutrition are desired for each of 3 districts in order to compare them. A sample size of 210 children is needed from which to draw valid estimates of nutritional status (i.e. 95 percent confidence in an estimate with +/- 10 percentage points in either direction) and infer the results to the larger population. Therefore, a sample of 210 is taken in each district to allow for separate estimates, at the defined level of confidence and precision, from each district. This results in a total sample size of 630, which means that the overall estimate (for all the districts combined) will be more precise (based on 630 instead of 210 units).

Example of Stratification of a Non-probability Sample

10 villages are to be included in a sample for conducting focus group discussions. Just before setting off for the villages, several staff members point out that having mixed gender focus group discussions will prevent women’s participation owing to cultural roles in public meetings. It is decided that focus group discussions must be held with men and women separately in order to allow each gender’s varying experiences regarding participation in the food-for-work (FFW) activity to be compared. Consequently, 2 focus group discussions (with a total of 20 participants) are held in each of the 10 villages.

Examples of the Stratification Factors that are listed in Indicators

	Results hierarchy	Performance indicators
Impact	Decreased drop-out rate and increased completion of primary education by girls	A. % of students dropping out by gender and grade (during the academic year) B. % of students dropping out between one grade and another by gender and grade (between grades) C. Numbers of girls completing grade 6 and grade 9
Outcome	Increased enrolment of female students by an average of 10% per year over 5 years	A. Number of girls enrolled at the beginning of each academic year
	Increased attendance by enrolled female students	A. % of students absent 3+ days/month by gender
Output	Dry take-home rations distributed (on average per year) to female students who are enrolled and attending classes	A. Number of rations distributed to girl students per semester/per academic year

Comparison Groups: why not to use Them

Introduction. The purpose of this section is to explain the rationale for using comparison groups and why it is inappropriate for the majority of WFP operations to do so.

What is a Comparison Group

A comparison group is a group of individuals who are not exposed to a WFP operation, but who share characteristics similar to those of the operation's target group.

Comparison Groups: why not to use Them

The use of comparison groups makes it easier to attribute observed change over time to a particular phenomenon. When assessing the impact of WFP operations, the phenomenon of interest is the operation itself. Most WFP operations assess the impact of operations through comparing indicators before and after the operation, inferring that the change found is to some undeterminable degree related to the WFP operation.

The use of comparison groups adds strength to the assertion that WFP operations have, indeed, caused the change observed by comparing, before and after the operation, 1 or more groups that were exposed to the operation and 1 or more groups that were not. This allows evaluators to control for the affect of external or extraneous factors that influence the indicator of interest and to assess and quantify more thoroughly the cause and effect relationship between the WFP operation and the outcome of interest.

Identifying an 'exact' matching comparison group in terms of all the variables that are potentially related to the variable of interest may seem to be an impossible task. However, the exposed and unexposed groups must share at least an approximately equal condition for the primary variable of interest.

Therefore, the main issue that makes the use of comparison groups inappropriate when evaluating the majority of WFP operations is that WFP's mandate, particularly in PRROs and EMOPs, is to enact full-coverage operations in areas where populations are hungry or considered hungry poor. In other words, comparison groups that are not exposed to the operation and that share the same condition for the variable of interest do not readily exist. Furthermore, it is inappropriate and unethical to withhold food aid simply in order to improve the design of an evaluation study.

However, where operations are not full-coverage owing to resource constraints, the use of comparison groups may be warranted. A final consideration against using a study design that incorporates comparison or control groups is the increased cost of doing so (e.g. the increase in sample size). Often cost is the prohibitive factor to the use of comparison groups, even when the ethics of doing so are not in question.

An Example of the Difference between using and not using Comparison Groups

The formulas presented below clearly identify the difference between using and not using comparison groups. When using comparison groups the additional considerations of 'extraneous factors' allow for a shift in conclusions from change associated (i.e. when no comparison groups have been used) with WFP operations to change attributed to WFP operations.

Before and after Evaluation Strategy with no Comparison Group

Pre-programme condition for outcome and impact indicators	—	Post (or mid-term)-programme condition for outcome and impact indicators	=	Change over time associated with WFP operations
-----------------------------------------------------------	----------	--------------------------------------------------------------------------	----------	-------------------------------------------------

Before and after Evaluation Strategy with Comparison Group

Pre-programme condition for outcome and impact indicators	—	Post (or mid-term)-programme condition for outcome and impact indicators	+ / -	Effect of extraneous factors	=	Change over time attributable to WFP operations
-----------------------------------------------------------	----------	--------------------------------------------------------------------------	--------------	------------------------------	----------	-------------------------------------------------

Preparing the Baseline Work Plan and Budget

Introduction. The purpose of this section is to describe what should be included in a baseline study work plan and budget.

Proposed Content for Baseline Study Work Plan and Budget

Once the design and methodological issues related to the baseline study have been decided, they should be summarised in a study plan. This is critical for comparative purposes as it ensures that subsequent studies (mid-term and final evaluations) can duplicate the methodology used. The costs associated with the baseline study should also be detailed in a budget.

Summary (should include a timetable)
Background and purpose of study <ul style="list-style-type: none"> ● Description of operation design and target beneficiaries ● The objective of the study - list of indicators from the logical framework that the baseline study measures ● Review of existing data sources
Data collection <ul style="list-style-type: none"> ● Defined units of study ● Proposed use of secondary data ● Proposed primary data collection methods and techniques ● Sampling description
Design <ul style="list-style-type: none"> ● Questionnaire or checklist ● Arrangements for pre-testing
Fieldwork <ul style="list-style-type: none"> ● The fieldwork team ● Training required for enumerators ● Timetable for fieldwork ● Description of quality controls/supervision in the field (e.g. supervision and arrangements for data checking and filing)
Data processing and analysis <ul style="list-style-type: none"> ● Arrangements for data cleaning ● Arrangements for data entry and processing ● Proposed framework for analysis - proposed data tables, indicator calculations and stratification/disaggregation of data ● Training required in data management and analysis
Reporting and feedback <ul style="list-style-type: none"> ● Proposed format of study report (preliminary table of contents) ● Arrangements for presentation/dissemination of findings
Annexes <ul style="list-style-type: none"> ● Budget ● Operation design document

Although it is not possible to provide clear guidance on how much to spend on collecting primary baseline data, the following list of items may need to be incorporated into the budget for baseline data collection.

Budget heading	List of possible budget items
Personnel	<ul style="list-style-type: none"> ● Staff salaries and allowances <ul style="list-style-type: none"> ● contracted consultants ● field staff involved in data collection ● field supervisors ● data entry and processing staff ● drivers
Training	<ul style="list-style-type: none"> ● Hire of training venue and accommodation ● Transport to field sites for practical training exercises ● Hire of training equipment ● Food and beverages ● Training materials and stationery
Transport	<ul style="list-style-type: none"> ● Vehicle operating expenses ● Allowances for maintenance of motorcycles or bicycles purchased specifically for the study
Equipment	<ul style="list-style-type: none"> ● Measuring equipment (e.g. weighing scales, measuring tapes) ● Field staff equipment (e.g. clipboards, calculators) ● Computer equipment (e.g. hardware, software, consumables) ● Transport (e.g. bicycles or motorcycles)
Stationery	<ul style="list-style-type: none"> ● Paper for questionnaires or checklists (including spares) ● Manuals (e.g. for data collection, data entry) ● Reporting pro formas for monitoring of data collection ● Pens, pencils, sharpeners, erasers, rulers, etc. ● Report production
Publicity	<ul style="list-style-type: none"> ● Posters ● Leaflets ● Hire of rooms for meetings ● Radio announcements ● Dissemination of study results (e.g. through a workshop)
Contingency	<ul style="list-style-type: none"> ● Allow for unexpected problems and delays during field-work

Module Summary

What has been covered in this module?

This module explained what a baseline study is and how it is related to the mid-term and terminal evaluations. It described when to conduct a baseline study and how to ensure that the outcome and impact indicators on which the study will collect data are clearly stated and, not least, in line with the operation design as outlined in the Logical Framework.

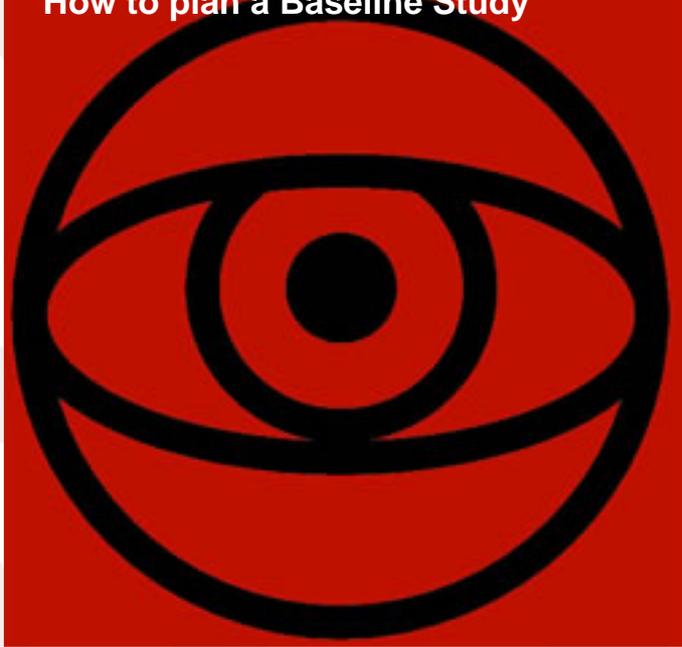
The module defined key concepts related to baselines and provided guidance on the use of primary or secondary data, sampling techniques, how to define the unit(s) of study, whether to use comparison groups, etc. In addition, it described what a WFP baseline study work plan and budget should look like.

What additional resources are available?

For further information the following modules and resources might be useful:

- How to Plan an Evaluation
- How to manage an Evaluation and disseminate its Results
- How to plan and undertake a Self-evaluation
- Choosing Methods and Tools for Data Collection
- Going to the Field to collect Monitoring and Evaluation Data
- How to consolidate, process and analyse Qualitative and Quantitative Data
- Reporting on M&E Data and Information for Development Programmes
- Reporting on M&E Data and Information for EMOPs and PRROs

How to plan a Baseline Study



United Nations
World Food Programme
Office of Evaluation and Monitoring

Via Cesare Giulio Viola, 68/70 - 00148
Rome, Italy

Web Site: www.wfp.org
E-mail: wfpinfo@wfp.org
Tel: +39 06 65131