World Vision

# Baseline Field Guide

June 2018

Our vision for every child... Life in all its fullness. Our prayer for every heart... The will to make it so.

# TABLE OF CONTENTS

# Abbreviations

| | |
|---|---|
| ANOVA | Analysis of Variance |
| AP | Area Programme |
| BDD | Baseline Design Document |
| BWS | Question Code: Module B of the caregiver survey with questions on Water and Sanitation |
| CESP | Community Empowerment and Sponsorship Planning |
| CGS/CG | Caregiver Survey |
| CWB | Child Well-Being |
| DAP | Development Asset Profile |
| DEFT | Design Effect |
| DHS | Demographic Health Survey |
| DME | Design Monitoring and Evaluation |
| DPA | Development Programme Approach |
| EGRA | Early Grade Reading Assessment |
| ENA | Emergency Nutrition Assessment |
| F&D | Faith and Development |
| FGD | Focus Group Discussion |
| FLAT | Functional Literacy Assessment Tool |
| GC | Global Centre |
| HH | Household |
| KII | Key Informant Interview |
| LB | Literacy Boost |
| LEAP | Learning through Evaluation, Accountability and Planning |
| M&E | Monitoring and Evaluation |
| MoE | Ministry of Education |
| MICS | UNICEF Multiple Indicator Cluster Survey |
| MVC | Most Vulnerable Children |
| ND | National Director |
| NO | National Office |
| ODK | Open Data Kit |
| P&C | People and culture |
| PST | Programme Support Team |
| RO | Regional Office |
| ROI | Return on investment |
| SMART | Standardized Monitoring and Assessment of Relief and Transitions |
| SO | Support Office |
| STAR | School-based Test About Reading |
| SST | Strategy Support Team |
| TA | Technical Approach |
| TOR | Terms of Reference |
| TP | Technical Programme |
| TPM | Technical Programme Manager |
| UNICEF | United Nations International Children's Emergency Fund |
| WASH | Water, Sanitation and Hygiene |
| WV | World Vision |
| YHBS | Youth Healthy Behavior Survey |

# INTRODUCTION

**Baselines: definition, purpose and value** A baseline is a study that collects and analyses qualitative and quantitative data and uses the results to establish the current status of important child wellbeing (CWB) indicators **before** implementation of a project or programme. Collecting baseline information before programmes begin (or within the first year of implementation) helps World Vision and partners in the following ways:

1. It helps national offices (NOs) to better understand the current context, validate their strategic choices and inform investment decisions.
2. It helps NOs to set benchmarks against which the impact of programmes is assessed in the future.
3. Baseline information helps World Vision (WV) staff and partners to strengthen their programme monitoring and evaluation.

**Baseline for LEAP 3:** LEAP 3 requires baseline data to be collected for all WV-supported projects/programmes within the first year of implementation. This standard applies to all programmes irrespective of their funding source (sponsorship, private non-sponsorship and grants)[1]. Additionally, LEAP 3 recommends that baseline data for programmes should only be collected: (a) from the geographical territory (usually communities) where WV intends to implement child-focused programmes - technical programmes (TPs) and area programmes (APs) and; (b) on indicators that will measure the progress and success of interventions within priority sectors that national offices have invested in (as outlined in national office strategies and technical programmes).

**How the field guide is organised:** This field guide has been developed to guide World Vision staff and partners in the planning and carrying out of programme baselines. In order to simplify the key concepts and principles for use, the field guide is organised around three principal phases of managing baselines: planning, execution, and data analysis and use. The field guide illustrates key steps to follow as WV staff and partners plan, implement and use baseline information

**Pre-conditions:** Before national offices plan to conduct baselines, they should ensure that programme designs are 'completed.' This means that the TP design document has been agreed by the funding office, implementing partners and has been approved by donors (for grant funded programmes), by National Directors (for TPs) and by Support Offices (for PNS projects). Because major donors often have set baseline standards and business processes, more attention on this topic will be given to technical programmes. (For more guidance on TP preparation and finalisation, please refer to the LEAP 3 TP design guidance, template and quality assurance framework).

For TPs, LEAP 3 requires that monitoring and evaluation (M&E) plans are jointly developed/agreed upon by WV NO staff, programme support teams (PSTs) and other partners. M&E plans should clearly show the key indicators WV and partners will track at all levels of the TP (and by APs contributing to the specific TP). This will enable national offices to assess the implementation of strategy and report on progress made towards important CWB targets.

[1] This field guide provides recommended principles and standards for conducting baselines for APs and TPs. Whereas, these principles and standards would apply to grants, it is important to note that donors (especially governments) normally prescribe their own standards and processes which must be followed while collecting baseline information for grants.

# PHASE 1: BASELINE PLANNING PHASE

Effective planning for baselines uses 11 steps which are outlined below.

## Phase 1 Step 1: Organise the baseline planning team

Effective planning, implementation and use of baselines requires the coordinated efforts of many stakeholders (both within and outside WV). To enhance team coordination and effectiveness, this field guide proposes that each baseline effort should be coordinated by a team of 8 -12 persons. These should include:

- **A team leader** – ideally a senior WV national office DME manager with significant experience in planning, execution and accountability for baselines and evaluations. The team leader is accountable for the overall planning and coordination of the baseline process (and will ensure consistency and alignment with the standards outlined in this field guide).
- **The team leader is supported by a team of 10 to 12 people**. These should include: (a) a senior NO operations/programmes leader – who provides leadership direction and resources for the baseline; (b) 2 to 3 technical programme managers (TPMs) – who ensure baseline purposes/indicators, methodologies are technically sound (TPMs are accountable for meeting all technical standards during baseline planning and management); (c) 3 AP managers – who will guide the implementation of baselines in the primary focus areas; (d) 1 to 2 representatives from the NO's PST and (e) representatives from the NO with experience in finance, faith and development, and people and culture (P&C). Where the local context permits, NOs should consider including representatives from government and key local partners (including faith communities and faith-based organizations) who can assist WV staff to mobilise local participation and ownership of the baseline process.

Before the baseline team starts planning, it is important that all members of the team read the baseline field guide as well as the design document of the programmes they intend to baseline. During the course of planning and implementing a baseline, baseline teams are encouraged to bring in technical expertise from other parts of the WV partnership (i.e. DME and technical specialists located at regional offices, support offices and the global centre). LEAP 3 recognises that during the process of planning and executing baselines, national offices may also need technical assistance from external consultants. When this need arises, field offices and their PSTs must ensure that engagements with consultants are guided by clear terms of reference. For more guidance, baseline teams should consult WVI DME/DPA guidance on developing consultant TORs (and managing their performance).

## Phase I Step 2: Agree on baseline scope, purpose and objectives

Planning and collecting baseline information requires significant investment of financial, human and other resources. Like any other investment, baselines should be planned and carried out in a manner that maximises the return on that investment (ROI). This requires programme baselines to be strategically focused and guided by clear objectives and questions that will help programme staff maintain focus on baseline objectives during all crucial stages of the baseline survey process. Collecting baseline information helps WV and partners to:

a. set benchmarks against which programme impact can be monitored and evaluated
b. validate and strengthen programme targeting and technical approaches
c. strengthen their understanding of existing needs/opportunities.

All the purposes and objectives of a baseline study should be in line with these aims.

Sometimes a national office will have additional baseline purposes and objectives. For instance, it may need to advocate for child-related government policy and programming at the sub-national level (say at the county level) because the country has just passed legislation to devolve governmental functions and decisions down to the county level. For such an office to be effective in its advocacy work with county

governments, it would need to use its TP baselines to collect additional county-level data. It is important that a NO spells out this additional baseline objective because it may impact on the methodology, specifically on sampling and sample size.

When planning and implementing a baseline study, it is important to remember that all the key phases and steps are inter-related. The purpose, objectives and scope set at the planning stage will influence key decisions and choices regarding: (a) what data should be collected; (b) what methods should be used to collect that data; (c) how the data should be analysed. Baseline teams must invest sufficient effort to ensure that this stage of baseline planning is done well. LEAP 3 recommends that the baseline team and stakeholders should continue to reflect (and where necessary refine) baseline purpose and objectives in order to facilitate efficient execution of data collection and analysis.

**Scope of the baseline**

Over the years, WV has gained substantial experience in planning and conducting baselines at project and AP levels. These baselines have been relatively easy to implement. This is because they have often targeted AP projects with limited scope implemented in localised geographical territories. However, LEAP 3 baselines require NOs to carefully analyse their portfolios and determine which programmes should be part of the baseline's scope, and why. Before suggesting a methodical approach, here are some guiding principles that need to be considered by NOs as they plan for LEAP 3 baselines.

a) TPs are the primary unit of planning and analysis under LEAP 3. Therefore, key questions regarding the baseline scope and sampling should first be considered at TP level (and subsequently at AP level). The required sample size for TP baselines should be determined at TP level while ensuring that samples at AP level are adequate to enable meaningful analysis and conclusions at AP level.

b) In most national offices, TPs include activities that will be implemented by APs, (based on grants) and national-level activities such as advocacy interventions. TP interventions at both levels should be adequately baselined. With the exception of APs that will transition during the current NO strategy cycle, all APs contributing to specific TPs should be included in baselining (and subsequently evaluation) of the TPs they contribute to.

c) APs contribute to TPs via implementation and scaling-up of project models that are most appropriate for their contexts. Only project models actually implemented by the AP (or that will be implemented) will be baselined

d) In each AP, relevant components of TPs are implemented in specific primary focus areas. NOs should conduct detailed planning to identify the primary focus areas where different components of TPs will be implemented during the current cycle of NO strategies. Baseline data should then be collected from these primary focus areas. Information from non-primary focus areas should only be collected during situations where the NO/partners choose to establish control groups[2].

e) LEAP 3 encourages national offices to consolidate their TP baselines and conduct as few baseline surveys as necessary. However, consolidating baseline needs should be done carefully to ensure the baseline needs of all TPs are addressed, and measurement standards developed for different project models are met.

**Necessary Steps**

In line with the five principles above, the following steps are recommended to assist national offices and partners determine the relevant scope for their LEAP 3 baselines.

**A. Mapping of TPs and key indicators:** As part of local adaptation of TPs, NOs should carefully analyse their programme portfolios and identify APs that will contribute to specific TPs. Typically, these would be APs where WV staff and partners have agreed to focus on development problems which the specific TP (e.g. health and nutrition) is designed to address. For example, during this mapping process, if a national office has 30 APs and 3 TPs (health/nutrition, livelihoods and education) it should be able to determine which of its 30 APs will contribute towards: the

[2] Programmes should note that there are standard protocols (including sample size considerations, data analysis issues) that need to be followed whenever a programme wishes to use control groups. NOs desiring to use control groups or other quasi-experimental designs in their baselines should seek guidance from regional and GC technical leaders.

health/nutrition TP, the livelihoods TP and the education TP respectively during the current strategy cycle. All APs contributing to different TPs should be included as potential candidates to participate in baselining of these TPs. There are three reasons why LEAP 3 recommends that all APs contributing to a specific TP should be included in the baselining of that TP.

- **APs remain an important model for delivering field ministry and need information from TP baselines to enhance their planning and management quality**. At the conceptual stage of LEAP 3, it was thought that NOs would be able to consolidate a group of APs to form an integrated Area Programme (with one design document, pooled funding and M&E system). However, consolidation of programmes is associated with many programmatic and financial risks that need to be carefully examined (through pilot tests). In the interim under LEAP 3, TPs are the basic unit of planning and analysis, APs will remain an important model for delivering field ministry. Including all APs which contribute to a specific TP in the baselining will give AP teams concrete information that helps them to better understand needs, vulnerabilities and opportunities that exist in their primary focus areas. This enables WV staff and partners to select the best project models that can be scaled up in order to address child well-being (CWB) priorities.
- **Information from TP baselines will help APs to strengthen their M&E systems**. TP baseline information will assist AP teams/partners to strengthen their M&E systems by helping them set appropriate performance targets, identify crucial milestones, and improve ongoing monitoring and reporting focused on strengthening programme impact. These will align with priorities identified by TP baselines.
- **Information from TP baselines will help NO leadership to enhance focus and investments in their APs.** By including all APs contributing to a specific TP in the baselining of that TP, national offices will gain clear evidence that helps them better understand differences in vulnerabilities, needs and progress among their APs. With this information, NO leadership teams will be able to make better decisions and guide investments towards the most crucial needs existing within and across APs. NOs will also be able to set performance targets and milestones for all NO strategy indicators across all their APs which contribute to those indicators. This will help identify and support APs with CWB gaps to address those gaps and achieve acceptable performance thresholds.

**B. Mapping of Project Models:** After identifying APs that will contribute towards different TPs, NOs/partners should carefully map-out project models that will be implemented by each AP contributing to a specific TP. This process (which is part of the local TP adaptation work) will enable NOs to identify which project models (and indicators) should be baselined and from which APs.

**C. Mapping of Geographical Area**: As part of their strategy development process, NOs should identify geographical areas where they will invest during the next 3-5 years of their current NO strategy cycle. These areas should include primary focus areas where WV currently implements programmes, as well as locations where NOs expect to grow (using their sponsorship, private and grant revenue streams). Baseline data should be collected from the geographical area where WV plans to implement its TPs during the current cycle of its NO strategy.

**D. Analyse AP life cycles (and eliminate transitioning APs)**: Only study units that were included in baselines should be considered for evaluation. This principle enables programme stakeholders to assess programme effectiveness by comparing the baseline and evaluation status of indicators of success. For this reason, it is very important for NOs to review their programme portfolios and identify APs that will transition/phase-out before TPs are evaluated. Such programmes should be excluded from baselining[3]. However, NOs should allow transitioning APs to contextualise and implement relevant TPs and the NO CESP priorities.

[3] Separate reviews/evaluations should be used for programmes that will transition before the TPs they contribute to are evaluated. Similarly separate baselines and reviews/evaluations may be organised for grants that may be started in between TP baselines and evaluations.

**E. Mapping of Target Beneficiaries:** NOs design and implement programmes that are intended to enhance the well-being of children, especially the most vulnerable. The WVI guidance followed by NOs during the process of developing their NO strategies requires field offices to think-through and identify the different populations who will be targeted by different TPs and project models. By identifying target beneficiaries, this will enable programme staff to ensure these beneficiaries are included in baselining of the relevant TP.

Different project models will require different sampling methods. Some require that baseline information for their key indicators is collected using population-based surveys. Others recommend that baseline information should be collected from project beneficiaries. Therefore, selecting geographical territories where TPs will be implemented enables NOs to include the potential beneficiaries of different TP interventions in integrated baseline processes. LEAP 3 recognises that in most cases, it is practically impossible for NOs to select all populations who will benefit from TPs. For example, at the beginning of strategy implementation, a national office may not be in position to project the number of schools that will participate in Literacy Boost/ Unlock Literacy interventions when the NO secures future grants. During these situations, LEAP 3 encourages NOs to implement 'rolling baselines' – collecting baseline information as and when project beneficiaries are selected. This baseline information is then used to establish strong M&E systems to enable stakeholders to track and report on programme impact[4].

**F. Develop a preliminary sampling frame:** In statistics, a sampling frame refers to all units and elements from which a representative sample could be selected for a survey/research. Each NO should therefore develop a detailed list of possible sample units for each AP - based on the different types of indicators to be measured during baseline. This list is informed by the type of TPs the NO plans to implement, and the APs that are contributing to different TPs (identified following the steps A-E above).

## Phase 1 Step 3: Review technical programme outcomes, indicators, indicator definitions and socio-economic information

**TP outcomes, indicators and indicator definitions:** TP M&E plans should clearly show the key indicators to be tracked to enable NOs to effectively assess strategy implementation and report on progress they are making towards realising important CWB targets. During the process of selecting indicators, NOs are encouraged to use industry-tested indicators standardised by WV and contained in the Compendium of Indicators[5]. Once TP M&E plans have been completed, NOs can assess and identify outcome indicators of the various TPs that need baseline information.

To begin this step, the baseline team should retrieve all TP-related documents. For each TP, the team should create an excel file named Indicator Mapping Table (name of TP) with 5 columns and illustrated in the table below. Baseline teams should map-out all TP indicators using this spreadsheet to ensure that all indicators are correctly defined (in accordance with the compendium of indicators). Any mistakes in indicators should be corrected at this stage.

[4] Rolling baselines can be used to strengthen programme baseline and M&E by collecting baseline information progressively as programmes are initiated/expanded. For example, if a national office is planning to implement Literacy Boost activities in 5 provinces during a 5 year period and plans to start implementation in one province during the first year and expand to other provinces during year 2, baseline information from province one can be collected during year 1 and baseline information from other provinces collected during year 2 after communities/schools have been selected.
[5] Indicators in the Compendium should be compared against other key indicators, i.e. Demographic Health Survey (DHS) to allow for consistency within the country. In such cases, NOs may want to use additional indicators. The process of selecting such indicators should be guided by technical specialists in NOs and ROs. Additionally, it is recommended that programme staff follow data measurement protocols (reference sheets, data collection tools) recommended for each indicator contained in the compendium of indicators.

*Example 1: Indicator mapping table (WASH TP)*

| Outcome | Indicator | Indicator definition | Target group | Indicator computation |
|---|---|---|---|---|
| **Improve household (HH) inclusive access to sanitation facilities and hygiene practices by 25% by 2020.** | Proportion of households with access to improved and safe sanitation facilities for defecation. | HHs using an improved sanitation facility - typically a latrine or toilet for defecation. An improved sanitation facility is one that deals hygienically with human excreta. This includes VIP latrines, pit latrine with slab and composting toilets. | Households | Indicator = number of HHs with access to improved sanitation facilities divided by total number of HHs. |
| | Proportion of households with parents or caregivers with appropriate hand-washing behavior. | Context is appropriate behavior change methods. Hand-washing at critical times and appropriate methods (use running water and soap/ash). | Households | Indicator = number of HHs with parents or caregivers practicing appropriate hand-washing divided by total number of HHs. |

More rows should be added to the spreadsheet if the TP has multiple outcomes and indicators.

**Socio-economic information.** Programme baseline measurements should collect information about priority socio-economic issues that will enable programme teams and partners to better understand their target community and strengthen the feasibility and appropriateness of proposed interventions. In practice most programmes (TPs and APs) are designed using secondary data sources and/or quick rapid assessments. These may not provide adequate information about the resources, vulnerabilities and development priorities of target populations.

Here is an example of a TP development process to illustrate the importance of using baselines to collect important socio-economic information. LEAP guidance and standards require that NOs should develop their strategies before they develop TPs (which outline the steps the NO will take to achieve its strategic objectives). During the process of developing NO strategies (and subsequent TPs), NOs will conduct landscape analysis based on available secondary data to identify the most significant CWB problems. The NO will focus on these and determine the most appropriate technical approaches for optimum impact and effectiveness. The use of secondary data for strategic planning usually has some limitations. For example, UNICEF country reports and demographic health survey (DHS) data may be helpful in highlighting the prevalence of child malnutrition in a country, but those reports may not adequately show the extent of malnutrition among the primary focus areas where WV invests (or intends to invest). Similarly, such reports may not contain important information about demographics, households, community assets, infrastructure and the spiritual/faith landscape etc. which WV staff/partners need to know in order to strengthen the focus, relevance and effectiveness of their programmes.

To overcome this challenge, LEAP 3 recommends that programmes should use baseline surveys to collect priority socio-economic information (both quantitative and qualitative) to help programme staff/partners better understand their target communities and available community resources. This will enable them to validate/adjust their programme focus and targeting approaches. Examples of socio-economic information that could be collected by baselines include:

- Important demographic data (population sizes, settlement patterns, etc.)
- Community resources/infrastructure - transport systems (roads, quality and accessibility); access/quality of water resources; health and education facilities; access to markets; etc.) This kind of information should be collected through the Development Programme Approach (DPA) processes which APs engage in before and during the baseline.

- Household specific information (family sizes/composition, economic activities, ownership of priority assets, etc.)[6]

Socio-economic information can be broadened to explore important areas that are vital to WV's development programmes approach. These include:

- **Faith and development issues** – Being a Christian organisation, WV has made a commitment to ensure that faith issues are properly understood and addressed in all areas where WV works. During baselining, information should be collected that helps programme staff to effectively: (a) understand faith/beliefs that exist in communities and how different faiths influence development choices; (b) explore existing faith-based organisations/faith leaders, and the roles they play (and could play) in facilitating transformational development and; (c) assess the current state (strengths, opportunities, challenges) of spiritual nurture for children. Exploring faith and development (F&D) issues at baseline would help NOs gain deeper understanding about what is happening in communities, the theological interpretations of what is happening in communities, and the extent to which the gospels are being upheld, etc.[7].
- **Progress out of poverty** – WV's development approach aims at helping poor and vulnerable communities to address the root causes of poverty and achieve sustainable reduction in poverty levels. For this reason, programme baselines should collect information that enables WV staff and partners gain a better understanding of the drivers and intensity of poverty among WV-supported communities. A standardised methodology called progress out of poverty index (PPI) has been developed and tested by Vision Fund and some NOs and is highly recommended for all NOs that want to assess poverty levels and the progress their communities are making out of poverty. (More guidance on PPI can be accessed at https://www.povertyindex.org/).
- **Information on crucial advocacy issues -** Similarly, socio-economic baseline information can be broadened to help NOs collect country-level information on interventions implemented at national level (such as advocacy and policy engagement processes). This is important because typical TPs include national level advocacy priorities whose progress should be tracked following sound M&E processes (including identifying and baselining indicators, setting performance targets; etc.).

**Caution:** When considering what socio-economic information to include in programme baselines, programme staff and partners should only prioritise information that helps them to better understand the community and strengthen their programme designs, delivery and performance management processes. The type and extent of socio-economic information to include in programme baselines should be determined by the scope of TPs and project models that will be implemented by TPs. Such decisions should be strongly guided by NO technical specialists.

[6] Socioeconomic information should be collected from target populations (i.e. children, youth, caregivers) who are selected to participate in the baseline survey

[7] For more information regarding faith and development issues which should be explored at strategy/TP development and baseline time, NOs are encouraged to refer to the "Spiritual Landscape Analysis Tool" and contact their regional F&D technical leaders.

## Phase 1 Step 4: Determine appropriate data collection methods

Assembling baseline information does not always require programme staff/partners to collect primary data. In some cases, good quality secondary data may exist that should be assessed before a decision is made to collect primary data. For example, if an area programme is designing the expansion of a health project that was started and baselined one year ago and covered all intervention areas and key indicators, there may be no need to collect primary data for the expansion phase of its project. However, in most cases, good quality secondary data that is timely and covers all areas of interest and key indicators does not exist. For this reason, programmes are encouraged to collect primary data for their baselines and only use secondary data sources to triangulate the key findings of their baseline surveys. Secondary data sources that could be used to triangulate baseline findings include: demographic and health surveys (DHS), population censuses, UN reports (such as UNICEF health/nutrition reports and WFP vulnerability reports), analytical reports undertaken by academic, research, and/or development institutions and evaluations/baselines implemented by WV or other organizations.

After the decision to collect primary data has been made, the next step is to specify the methods and tools to collect data. There are a lot of resources (generic tools and guides) readily available within WV which NOs can adapt/contextualize and use as appropriate. The most commonly used survey tools for quantitative data are listed in the following table.

| Tool | Description |
|---|---|
| **Caregiver Survey Questionnaire (CGS)** | Collects data from caregivers and/or household heads in a randomly selected sample of household survey. |
| **Youth Healthy Behavior Survey (YHBS)** | Collects data from youth aged between 12 -18 years in a randomly selected sample of household survey. |
| **School-based Test About Reading (STAR)** | Provides critical information about children's foundational reading ability. STAR replaces Functional Literacy Assessment (FLAT) as the recommended tool for assessing reading. It measures reading skills for Grade 3 students who are the focus of WV's literacy programming. |
| **Development Asset Profile (DAP)** | Measures indicators around the wellbeing of youth |
| **Anthropometric survey (can be combined with caregiver survey)** | Measures nutrition indicators. |

The most common tools for qualitative data collection are focus group discussion guides and key informant interview questionnaires. Such tools (generic ones) can be easily developed by NOs for most of the indicators that require qualitative data. After the baseline team has identified the appropriate data collection tools, it should update the indicator mapping spreadsheet to highlight the tools which will be used to collect baseline data for each priority indicator (shown in example 2 below).

*Example 2: Updated Indicator Mapping Table (WASH)*

| Outcome | Indicator | Indicator definition | Target group | Indicator computation | Data requirements | | Source of secondary data | Measurement method for primary data | Tool |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Primary Data | Secondary Data | | | |
| **Improve HH inclusive access to sanitation facilities & hygienic practices by 25% by 2020.** | Proportion of households with access to improved and safe sanitation facilities for defecation. | HHs using an improved sanitation facility - typically a latrine or toilet for defecation. An improved sanitation facility is one that deals hygienically with human excreta including VIP latrines, pit latrine with slab and composting toilets. | HHs | Indicator = number of HHs with access to improved sanitation facilities divided by total number of HHs | Yes | No | Not applicable | Survey | WV Caregiver survey |
| | Proportion of households with parents or caregivers with appropriate hand-washing behavior | Context is appropriate behavior change methods. Hand-washing at critical times and appropriate methods (use running water and soap/ash) | HHs | Indicator = number of HHs with parents or caregivers practicing appropriate hand-washing divided by total number of HHs | Yes | No | Evaluation and AP annual reports; | Survey | WV Caregiver survey |

**Developing Survey Questions:** After the baseline team has identified the data gathering methods for all indicators that require primary data collection, the next step is to generate the questions they will need to ask in order to collect the appropriate data for each indicator. During the process of generating survey questions, baseline teams should apply the following tips.

- Ensure the information required for each indicator is well understood by the survey team.
- Consider the anticipated analysis for each indicator and determine what additional information is required to support the analysis.
- Select questions from the appropriate generic questionnaire for each indicator and revise the question as appropriate.
- If some questions are not available in the generic questionnaire, create questions in consultation with the appropriate technical staff in the NO, RO and Global Centre (GC).
- Apply appropriate contextualisation and use consistent coding for each question on the questionnaire.
- Make sure the indicator definition (ID) section includes all the relevant information such as HH#, date of survey, enumerator and supervisor codes, AP name, cluster# and name.
- Create an indicator-questionnaire table to check if all indicators and required additional information have been considered in the questionnaire design (see example below)

*Example 3: WASH TP indicator-questionnaire table*

| Outcome | Indicator | Indicator Definition | Tool | Question |
|---|---|---|---|---|
| **Improve HH inclusive access to sanitation facilities and hygienic practices by 25% by 2020.** | Proportion of households with access to improved and safe sanitation facilities for defecation. | HHs using an improved sanitation facility typically a latrine or toilet for defecation. An improved sanitation facility is one that deals hygienically with human excreta including VIP latrines, pit latrine with slab and composting toilets. | Caregiver survey (CGS) | CGS question BWS12: Type of facility (What type of toilet facility do members of your household usually use? Flush or pour/flush toilet flushed to …?)<br><br>CGS question D601: Diarrhea cases? |
| | Proportion of households with parents or caregivers with appropriate hand-washing behavior. | Context is appropriate behavior change methods. Hand-washing at critical times and appropriate methods (use running water and soap/ash). | Caregiver survey. | CGS question BWS14c1: when are hands washed?<br>CGS question BWS15a: how do they wash hands?<br><br>CGS question D601: Diarrhea cases? |

For qualitative questions, baseline teams should consult GC sector DME specialists for guidance and tools.

## Phase I Step 5: Develop tentative analysis plan

The purpose of conducting a baseline study is to help programme staff/partners gain a better understanding of their target population, to validate proposed technical approaches and strengthen the targeting of interventions (ensuring they reach the most vulnerable). To achieve these objectives requires well-guided analysis of the data collected from each target population. Historically, WV baselines and evaluations have conducted basic descriptive analyses (that do not help programme staff make strategic decisions). Here is an example to illustrate this important point.

> During the process of developing its current strategy, WV Burundi conducted a landscape analysis that identified child malnutrition as the most strategic challenge inhibiting the well-being of Burundian children. To respond to this challenge, the NO designed a health and nutrition TP. During the baselining for this, WV Burundi needed to establish the current status of key child nutrition indicators (under 5s underweight, and stunting). However, if their baseline analysis had stopped at this level, WVB would not have learned very much about its target population of under 5 children. The only additional information they would have gained from the baseline would be to understand the variability that exists in key nutrition indicators across its different APs and primary focus areas.
>
> To maximise value from TP baselines (and validate its root cause analysis at strategy and TP levels), WV Burundi would need to conduct additional analysis that carefully examines key variables, their inter-relationships and the contribution they make to child malnutrition. Such analysis would need to be guided by clear analysis questions. Appropriate analysis questions can be generated by asking the following important question: What do we need to know about children under 5 in our communities? Below are examples of key analysis questions that can be asked by WV Burundi about child malnutrition in their communities:

- What is the current prevalence of child malnutrition in our APs and primary focus areas?
- What (if any) are the key differences in malnutrition levels between male and female children under 5?
- Is there a relationship between under 5 malnutrition and household income and food security income levels?
- Do parent/caregiver characteristics (education, age, occupation) affect child malnutrition levels?
- To what extent are orphans and children with disability vulnerable to child malnutrition?

During baseline planning, clear analysis questions should be set for all key populations that are targeted by a baseline study. Typically, these would include: children under 5 years, mothers/caregivers, youth, children of school going age, etc. Both quantitative and qualitative data will require analysis questions. Below are some examples of qualitative analysis questions that can be set for mothers/caregivers who are targeted for a WASH TP baseline.

- How do mothers/caregivers assess their current access to safe water and sanitation?
- In their opinion, what are the biggest challenges their communities face related to safe water and sanitation?
- What opportunities/solutions do they propose to address crucial safe water and sanitation gaps in their communities?

Once analysis questions have been developed, the next step is to select appropriate variables (or specific questions in the survey questionnaire) and corresponding analysis required to answer the analysis questions.

The table below provides an example of an analysis plan (quantitative and qualitative) for a WASH project:

*Example 4: WASH technical programme indicator-questionnaire table and analysis plan*

| Outcome | Indicator | Indicator Definition | Tool | Survey Question | Indicator Computation | Type of analysis required | Any statistical values required | Any statistical tests required | Type of software |
|---|---|---|---|---|---|---|---|---|---|
| **Improve HH inclusive access to sanitation facilities and hygienic practices by 25% by 2020.** | Proportion of HHs with access to improved and safe sanitation facilities for defecation. | HHs using an improved sanitation facility typically a latrine or toilet for defecation. … | Caregiver survey. | CGS question BWS12: Type of facility<br><br>CGS question D601: Diarrhea cases? | Recode BWS12 into BWS12impfacility, where:<br>1 = improved facility,<br>0 = unimproved facility | Prevalence of the indicator?<br><br>Crosstab (BWS12impfacility and D601) | Confidence Interval -95%. | chi-square (BWS12impfacility and D601) | SPSS |
| | Proportion of HHs with parents or caregivers with appropriate hand-washing behavior. | Context is appropriate behavior change methods. Hand washing at critical times and appropriate methods (use running water and soap/ash). | Caregiver survey. | CGS question BWS14c1: when do they wash hands?<br><br>CGS question BWS15a: how they wash hands?<br><br>CGS question D601: Diarrhea cases? | 1. Recode BWS14c1 to BSWS14wash, where 1 = washed hands, 0 = not wash hands<br>2. Recode BWS15a to BWS15appmtd, wherein 1 = appropriate method, 0 = not appropriate method<br>3. Compute BSW14wash + BSW15appmtd = BSWapprowash.<br>4. Recode BSWapprowash to BSWS14cwashbehavior, where 1 = appropriate washing (if BSWapprowash = 2), 0 = not appropriate washing (if BSWapprowash < 2) | Prevalence of the indicator<br><br>Crosstab (BSWS14cwashbehavior and D601) | Confidence interval -95%. | chi-square (BSWS14cwashbehavior and D601). | SPSS |

## Phase 1 Step 6: Agree on appropriate sampling methods

Data analysis planning for quantitatively measured indicators can only generate meaningful results if proper sampling is employed in gathering the data. A sample is a portion of a specific population that has certain features or characteristics. In the context of LEAP3 baseline, the target population is the entire number of caregivers, household and children, living within the geographical boundaries of APs. Sampling refers to the process of selecting study units (such as APs, households, schools, individuals) from a population and after studying the selected units to generalise results back to the whole population. Selecting an appropriate sample allows the baseline team to collect information at reasonable cost from a small group, which can be generalised to the population from which the sample was drawn.

However, if not carefully used, sampling can lead to biased estimations that may not truly represent the characteristics of the population from which the sample has been selected. In order for sampling to work, baseline teams should ensure the following two important conditions are met.

- **Adequacy** - the selected sample should be large enough to generate well-supported results which accurately reflect the key characteristics of the population from which the sample has been selected.
- **Representation** - the baseline team should carefully analyse the population of interest, identify any significant sub-groups and differences that exist amongst these subgroups, and ensure an adequate sample is selected from each important sub-group.

In order to ensure these conditions are met, the survey team needs to consider appropriate sampling methods and sample size determination techniques. There are two major types of sampling methods: probability and non-probability sampling.

**Probability sampling:** This technique gives every element in the target population the chance of being selected to participate in the survey. Probability sampling enables research teams to reduce sampling errors that are normally introduced (during non-probability sampling) when judgments are subjectively made by interviewers regarding who should be selected to participate in the interview.

There are four major types of probability sampling: simple random sampling, stratified sampling, systematic sampling, and cluster sampling. Simple random sampling is the most recognized probability sampling procedure. Stratified sampling offers a significant improvement on simple random sampling. Systematic sampling is probably the easiest one to use, and cluster sampling is most practical for large national surveys[8].

- **Simple random sampling** is a probability sampling procedure that gives every element in the target population, and each possible sample of a given size, an equal chance of being selected.
- **Stratified sampling** is a probability sampling procedure in which the target population is first separated into mutually exclusive, homogeneous segments (strata), with a simple random sample then selected from each segment. The samples selected from the various strata are then combined into a single sample.
- **Systematic sampling** (or interval random sampling) is a probability sampling procedure where random selection is made of the first element for the sample, and then subsequent elements are selected using a fixed or systematic interval until the desired sample size is reached.
- **Cluster sampling** is a probability sampling procedure in which elements of the population are randomly selected in naturally occurring groupings (clusters). A 'cluster' refers to an existing grouping of people. The heterogeneity of the cluster is central to a good cluster sample design. Ideally, differences within the cluster would be high, and differences between the clusters would be

[8] Sampling Essentials

low. However, the elements within each cluster should be as varied as in the target population. Ideally, the clusters would be small but not so small as to be homogeneous. The sampling units or clusters may be space-based, such as naturally occurring geographical or physical units (such as states, counties, census tracts, blocks, or buildings); organisation-based, such as school districts, schools, grade levels and classes, or telephone-based, such as area codes or exchanges of telephone numbers.

Cluster sampling is a commonly employed sampling technique in WV and worth further explanation. There are three types of cluster sampling based on the number of stages:

- **Single-stage sampling**: every element within a sampled cluster is included in the survey.
- **Two-stage sampl**ing: first, a sample of clusters is selected, and then samples of units from each are included in the survey.
- **Multi-stage sampling**: involves the repetition of two basic steps: listing and sampling; sampling procedures (simple random sampling, stratified sampling, or systematic sampling) at each stage may differ.

Cluster sampling is preferred when it is difficult to get a sampling frame of units. If the population is widely distributed geographically or occurs in natural clusters, cluster sampling would be the preferred option. Similarly, in situations where cluster sampling is convenient and reduces the cost compared to other sampling techniques, then a researcher may decide to use cluster sampling.

When compared to simple random sampling of the same sample size, cluster sampling is less representative of the population, introduces higher variances and more complexity in analysing data and interpreting results. In addition, cluster sampling yields larger sampling errors for samples of comparable size than other probability samples. Generally, cluster sampling has a larger design effect.

**Non-probability sampling:** This relies on a more subjective means of inferring something about the population of interest from a sample. There is no statistical theory, like that for probability sampling, to guide the use of non-probability samples. They can be assessed only through subjective evaluation. Failure to use probability techniques means that the survey estimates will be biased. In addition, the size of these biases and their likely under or over estimation will be unknown. Findings from non-probability samples cannot be generalized to the population from which the respective samples are taken. Non-probability sampling can generally only be used in qualitative assessments where the objective is to establish meaning rather than representative study outcomes.

## Phase 1 Step 7: Compute for sample sizes

In order for sampling to work, baseline teams must determine a sample size adequate for baseline results to generate conclusions that represent the key characteristics of the population. There are both statistical and non-statistical considerations that affect priorities and decisions in sampling. These are outlined below and should be used to influence the values used in calculating the sample size[9].

**Statistical Factors** – Items below are set by the baseline team, and have a direct effect on the sample size calculation.

- **Precision (Type I error):** The Z-score for probability of committing a type I-error ($Z_{\alpha/2}$) corresponds to the probability that an observed change, or difference between two point estimates (such as baseline versus evaluation) would not have occurred by chance. *The recommended default in the sample size calculator is $Z_{\alpha/2} = 1.96$ (assuming 2-sided test with $\alpha = .05$).* This is set at 5% and should not be changed.

[9] The ensuing discussion on sample size calculation, including definitions, statistical and non-statistical considerations, formulae, instructions, example, and the like, are available in the clickable link named, "LEAP 3 Baseline and Evaluation Sampling Guide" in WV Central's LEAP 3 page: https://www.wvcentral.org/community/pe/Pages/LEAP.aspx.

- **Power (Type 2 error):** The Z-score for statistical power ($Z_\beta$) is the strength of a sample size to detect a change between two point estimates (e.g. $P_2 - P_1$), if one occurred. *The recommended default in the sample size calculator is 80% statistical power, which is a $Z_\beta = 0.84$.* This is set at 80% and should not be changed.
- **Expected population proportion at baseline (P1):** This is obtained from secondary data. Estimates of $P_1$ in many countries for most key health and development outcome indicators can be obtained from previous surveys such as DHS or UNICEF Multiple Indicator Cluster Survey (MICS) reports. This can also come from results of recent and relevant evaluations conducted by the NO. *It should be set to 0.5 if unknown*
    - An estimate on the extreme ends (high or low) will require a smaller sample size than an estimate that falls in the middle (e.g. 50%). Use the resources available to make a good estimate from secondary data (e.g. DHS surveys, WV baselines), but if one does not exist, the programme will need to be conservative and use 50%.
- **Expected population proportion at follow-up (P2):** This is what the NO thinks is a realistic target based on the four factors outlined in the next bullet point below.. It is essential that the baseline team engages technical experts at various levels in WV to seek guidance on realistic P2 estimates. For health and nutrition programmes, the target setting tool (guidance and calculator) can be used to estimate targets[10].
- **Detectable difference:** The expected difference between two estimates ($P_2 - P_1$). This represents the minimum level/amount of change over a given period of time required to be considered statistically significant. Items that influence what is set as the expected population proportion to establish the minimum detectable change should be discussed closely with technical experts familiar with the intervention and include:
    - *Intervention Method:* interventions have a variable influence on the rate and degree of change. **Example**: direct distribution of vaccinations would see faster and larger effects on vaccination rates than a radio campaign.
    - *Fidelity of Intervention:* the degree to which an intervention is delivered as intended/according to guideline. **Example**: a project only implementing three of the five key components of an approach may see smaller and slower effects on the desired outcome.
    - *Resources to Implement:* allocated project budget and staff capacity. **Example:** a limited budget or budget reduction that does not allow the project to implement activities as planned, will reduce the expected level of change.
    - *Clinical Significance:* the practical importance of the effect - whether it has a real and noticeable effect on daily life. **Example:** The ability to detect a 3-5% change in the appropriate treatment of diarrhoea may not be very meaningful for the programme impact but that amount of change would be significant for stunting
- **Design effect (DEFT):** The sample size must be increased to account for some of the limitations with certain sampling methods such as 2-stage cluster sampling compared to a simple random sample. Estimates of DEFT for many countries for most key health outcome indicators can be obtained from the tables in Appendix B in DHS reports. *Where the DEFT is not known, the programme will need to be conservative and use 2, which doubles the sample size.*

[10] The purpose of Target Setting resources is to provide guidance to NOs for setting Health and Nutrition programme targets at outcome, coverage and reach levels. https://www.wvcentral.org/community/health/Pages/Target-Setting-Guidance.aspx

**Non-statistical factors**: The items below do not feature in the initial computation of the sample size, but are additional calculations or multipliers that may have an effect on the values selected for statistical items.

- **Prioritisation of information needs (baseline/evaluation objectives):** Baseline stakeholders (e.g. NO, PST, partners) will need to agree on the main objectives of their baseline/evaluations and prioritise information needs relative to budget or resources constraints.
    - **Priority indicators:** a technical programme will have several outcome indicators (10-15) and each will require a different sample size. Baseline stakeholders will need to agree which are the most important for strategy and programme design with regards to detectable difference, precision and level of analysis - in case a compromise needs to be made. ***Example:*** The stakeholders decide that the literacy indicator is more important to prioritise for the sample size than the enrolment indicator due to the programme design, which focuses on teacher training and learning materials.
    - **Disaggregation (strata):** if the data needs to be analysed by different population groups (e.g. gender, ethnicity, geographic area) called strata, then a sample needs to be calculated and drawn from each group separately. The number of strata/ disaggregation categories has significant effects on the sample size. ***Example:*** If the calculator estimates a sample size of 200 for the Development Asset Profile (DAP) and the programme needs to disaggregate by gender, then a sample of 200 girls and 200 boys making a total of 400 is needed.
    - **Level of analysis (strata):** this is very similar to disaggregation but is a newer concept for LEAP 3. Data can be analysed to represent different geographic / programme coverage levels: area, cluster or the entire technical programme. These are also strata. The stakeholders will need to discuss and agree the needs for detectable difference and disaggregation for each indicator for the different levels of analysis. A sample size will need to be calculated separately for each level and these parameters will directly affect the calculation. ***Example:*** For a nutrition programme, the stakeholders decide it is important to analyse levels of underweight children under 5 at an area level, but to analyse levels of exclusive breastfeeding across the entire technical programme. They also decide that the detectable change for underweight at an area level should be set at 10 percentage points while at the technical programme level it should be set at 5 percentage points for the sample size calculations.
- **Demographics / Sampling Frame:** The demographic make-up of a community or country will affect the likelihood of finding an individual from the target population in a household or school etc. It often means more households/schools will need to be visited to find enough individuals. ***Example:*** In an area with a low fertility rate, it may be hard to find households with children under 5 and even harder to sample sufficient numbers in the 0-6 month or 0-23 month age range.
- **Resources / budget:** The resources available including budget and personnel combined with the logistics of carrying out the survey, can limit sample size by restricting the number of enumerators that can be hired, distances that can be travelled and number of days for data collection. Baseline teams should ensure the determined sample size can be studied within the available resources.
- **Safety:** Insecurity may prohibit entry to certain locations or reduce the amount of time available for the survey team to complete their work. **Example:** An upcoming election may shorten the timeline, or an urban area may only allow work during certain hours of the day

Once the baseline team agrees on the options available, the proposed way forward should be submitted to the NO leadership for decision. SOs and PST members should also be consulted to ensure common understanding about the process and their expectations for future evaluations.

**Formulae to calculate the sample size.** In most sectors of programming that WV invests in globally, programmes normally use two categories of indicators. The first category consists of indicators that are expressed as proportions. Examples include: % of children U5 who are underweight; % of mothers and caregivers who practice safe hand-washing behavior; etc. The second category includes indicators that are

expressed in averages (such as mean household income levels; mean age of children who can read and write etc.). Though the principles to determine an appropriate sample size for each category of indicators are the same, the formulas to calculate the respective sample sizes are different. They are presented as follows:

**Formula 1 (Determining sample size needed to detect a statistically significant difference between proportions measured at pre and post intervention surveys)**

$$n \geq DEFT \frac{\left[Z_{\alpha/2}\sqrt{2P(1-P)} + Z_{\beta}\sqrt{P_2(1-P_2)+P_1(1-P_1)}\right]^2}{(P_2-P_1)^2}$$

**Where:**

n: Desired sample size

DEFT: Design effect

$Z_{\alpha/2}$: Z-score corresponding to the level of confidence with which it is desired to be able to conclude that an observed change of magnitude ($p_2$-$p_1$) would not have occurred by chance (i.e., *Probability of Type-1 error*) (unless otherwise noted, assume 2-sided test with $\alpha$ = 0.05; thus Z = 1.96).

$Z_{\beta}$: Z-score corresponding to the probability with which it is desired to be certain of being able to detect a change of magnitude ($p_2$ - $p_1$), if one occurred (i.e. *Power to avoid Type-2 error*) (unless otherwise noted, assume power of 80%; thus Z = 0.84).

$P_1$: This is the expected proportion that will be obtained from a baseline survey, or in a survey of particular survey domain. .

$P_2$: This is the expected proportion that will be obtained from a follow-up survey, or in a survey particular survey domain.

P: ($P_1$ + $P_2$) / 2

*Because all the variables required by this formula can readily be estimated, this formula is highly recommended for use in determining sample sizes of baselines and evaluations.*

**Formula 2: Determining sample size needed to detect a statistically significant difference between means measured at two data points (pre and post intervention surveys)**

$$n \geq DEFT \frac{2\sigma^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

**Where:**

n: Desired sample size

DEFT: Design effect

$Z_{\alpha/2}$: Z-score corresponding to the level of confidence with which it is desired to be able to conclude that an observed change of magnitude ($p_2$-$p_1$) would not have occurred by chance (i.e., *Probability of Type-1 error*) (unless otherwise noted, assume 2-sided test with $\alpha$ = 0.05; thus Z = 1.96).

$Z_{\beta}$: Z-score corresponding to the probability with which it is desired to be certain of being able to detect a change of magnitude ($p_2$ - $p_1$), if one occurred (i.e. *Power to avoid Type-2 error*) (unless otherwise noted, assume power of 80%; thus Z = 0.84).

$d^2$: The expected difference between the two means (e.g. $d^2$ for difference between mean 2.5 and 3.6 = 1.1).

$\sigma^2$: Population variance, which is the expected level of dispersion, or variability, about some mean in the population.

**Note:** Using this formula to compute sample sizes requires the baseline team to have estimates of population variance. If this is not available, it can be calculated from historical data captured through similar survey. If such historical data is not available, then it can be calculated from a pilot survey.

**Automated sample size calculator and indicator prioritisation tool:** WV has developed an automated spreadsheet that can assist its staff and partners to determine the appropriate sample size. A copy of this tool is contained in Annex 1. The calculator will automatically work out the sample size when the baseline team enters their P1 and P2 values. In order to estimate the sample size required at AP level, users need to enter the number of APs where each indicator will be measured and indicate whether the required level of analysis is at TP level or AP level. To effectively use the sample size calculator, the baseline team should consider the tips below.

**Tips for using sample size calculator**

**Step 1-** Open the LEAP 3 Sample Size Calculator. Then fill in the first three columns (column A, B and C) by entering the indicator, the tool and target group for all the TP outcome indicators that require baselining.

**Step 2 -** Enter estimates of Baseline Proportion (P1) and Evaluation Proportion (P2) for each indicator. First P1 should be properly estimated by using the best available information from secondary sources including reports of other similar surveys. Then enter the estimated target (P2) – which should be agreed upon by sector experts. As soon as P1 and P2 values are entered for one indicator, the respective sample size is generated. For indicators which require probability sampling (rather than cluster sampling), baseline teams should change the values of DEFT (Column D) from 2 to 1.

**Step 3 -** For each indicator, enter the number of APs (in Column M) where it will be measured. As soon as the number of APs is entered, the sample size per indicator for each AP is generated in the last column of the calculator.

**Step 4 -** Review the completed sample size estimations with all the stakeholders to determine the final common sample size by taking the non-statistical factors into consideration.

The final decision regarding which sample size to use and the level of analysis required, must be made by national offices/PSTs/other partners. This decision must be fully guided and endorsed by the technical specialists who guide strategic planning and programme quality in NOs. Discussing these decisions in PSTs helps NOs to draw upon technical expertise based in support offices and regional offices. As a minimum standard - the baseline design document and baseline report must make clear the justification for the sample size baseline stakeholders have agreed to use.

**Additional notes**

Some data collection tools such as Youth Healthy Behavior Survey (YHBS), DAP and STAR include guidelines on sample size determination. Baseline tools are encouraged to consider these guidelines during the process of sample size determination.

## Phase 1 Step 8: Agree on the final sample size

In order to determine the appropriate sample size for TP baselines, baseline teams will experiment with many indicators – which yield multiple sample sizes. However, only one of the sample sizes generated through this process should be used for the actual baseline survey. To enable them to select the most appropriate sample size for the baseline, WV staff and partners should consider the following tips.

1. Estimate the minimum sample size for each core TP indicator at the NO level using the sample size calculator. Ensure sustainability and faith-related indicators are also included since every TP should be implemented with sustainability in mind.
2. Identify the indicator requiring the highest sample size per TP - and if it is too high and difficult for NO to afford the cost of collecting data on that indicator, select the second highest and make this very clear in the BDD (to be discussed in a later section).
3. Divide this TP level minimum sample just between **those APs that will actually implement the TP t**o ensure the minimum TP level sample size is achieved. Note: including data from APs not implementing the specific TP will increase the chances of making the wrong conclusion (at evaluation) - that the TP wasn't effective in significantly improving the specific outcomes. If NOs decide to measure TP indicators even in APs that aren't implementing the TP, during evaluation, they should conduct analysis that compares change in ADPs implementing and those not implementing the TP.
4. Based on the final AP level sample size, estimate the number and percentage of TP indicators that each AP is powered to make the conclusion that the TP was not effective in the particular AP if change between baseline and evaluation is not statistically significant.

## Phase 1 Step 9: Prepare a baseline design document

**Process: The Baseline Design Document (BDD) is crafted by the national office and reviewed/endorsed by the PST/SST and Global Centre Ministry Strategy and Evidence (MSE) DME Team**

A baseline design document (BDD) is a plan that guides the journey of conducting a baseline. It is the primary document that helps WV staff, partners, and consultants to carefully think through all the crucial decisions and processes that should be followed in order to efficiently conduct baselines. It is therefore crucial for the BDD to clearly express the baseline purpose, scope and methodology in a language that can be easily understood by all key stakeholders. Because conducting a baseline measurement requires the collaborative engagement of many stakeholders, WV staff are encouraged to ensure that baseline BDDs are co-created by all key partners. In the case of TPs, this would require NO leadership, NO technical/operations staff working closely with their PSTs (and regional technical leaders) and key partners within the NO (i.e. ministry of health, ministry of education) to develop appropriate BDDs for the baselines needed. Important information to include in each component of a BDD is summarised in the table below.

**Key information that must be contained in each baseline design document**

| **Key BDD Topic** | | **Recommended content** |
|---|---|---|
| A. | Executive summary (1/2 pages) | Use this section to highlight crucial information about the planned baseline (purpose, objectives, expected results and how those results will be used to improve programme planning, delivery and effectiveness). |
| 1. | Background (1page) | Include important information such as key development issues in the country, NO strategic focus, and description of approved TPs that need to be baselined.. Background information should describe the baseline scope (including purpose/objectives, study questions, etc.). The background should also include information about the APs that will implement the TP being baselined (including the TP adaptation process). |
| 2. | Methodology (3-5 pages) | Highlight the types of quantitative and qualitative data that will be collected. Explain the techniques that will be employed to collect this data: such as data sources and respondents; sample sizes for both qualitative and quantitative components of the survey (and how that sample size was calculated). Describe the sampling strategy for how key study units will be selected (at TP level, APs, households, individual respondents, other study units such as schools for literacy baselining, etc.). The methodology should also highlight the different sampling techniques that will be employed (purposive sampling, stratified/cluster sampling and simple random sampling) and how these techniques will be used to produce a robust methodology. It is important to include limitations, ethics/integrity considerations, data quality assurance plan, baseline teams and training plans for these teams. Finally, the methodology should also highlight the data collection tools that will be used (including processes of development, pre-testing, revision, and finalisation). In cases where the national office plans to integrate components from different survey tools (e.g. DAP, STAR), the methodology should outline reasons for their choice of different tools, and how the different components of key tools will be integrated to enhance rigour. |
| 3. | Field data collection logistics plan | Describe how different stakeholders involved in the baseline survey will work together to plan and execute the baseline. Include all key processes that will be followed to collect baseline data (including composition, training of baseline teams; ethics and integrity during data collection), key stakeholders/roles, key milestones/deadlines, field transportation and schedules for training data collection teams. To develop a good logistics plan, WV staff and partners are encouraged to think through all key processes, and develop appropriate |

| | | |
|---|---|---|
| | | milestones and timelines for each key step. |
| 4. | Data management (analysis) plan | Outlines steps that will be taken by the baseline team to ensure collected data is appropriately handled and processed (including the transcription of qualitative data, the coding/entry/analysis of quantitative data and the storage of data/management of reports after the end of the baselining process. Describe the software that will be used to analyse both qualitative and quantitative data and the different types of analysis that will be conducted (e.g. frequency tabulations, cross-tabulations, t-tests and ANOVA) for various TPs and sustainability indicators. Note the types of variables that will be analysed (nominal, dichotomous, ordinal, interval, ratios) for TPs and the data interpretation plan. |
| 5. | Communication and utilisation plan | Describe the steps the baseline team will take to compile baseline reports and other key dissemination products. To complete this section, baseline teams should carefully think through different baseline audiences, the interest each audience has in baseline findings and develop appropriate communication products/mechanisms to effectively share baseline findings with different stakeholders. Describe how the principal baseline report will be compiled and its key components (executive summary, background, methodology, limitations, findings, lessons learned, conclusions/recommendations, appendices, etc.). Programme teams/partners are encouraged to innovate and employ creative (and cost-effective ways) to share highlights of their baseline findings. |
| 6 | Itemised budget | Present a comprehensive investment plan for the baseline costs related to: (a) personnel – including costs that will be spent on WV staff, consultants, data collection teams, etc.; (b) development of data collection tools/necessary technology; (c) equipment such as weighing scales for anthropometric measurements; (d) transportation – national level/air/field costs; (e) data entry/analysis costs. |

The BDD draft should be reviewed and endorsed by the PST/SST and GC and be approved by the National Director. Because of the need to pre-test the baseline design, and the number of other tasks related to the LEAP 3 process, review, endorsement and approval should not take more than a couple of weeks.

## Phase 1 Step 10: Develop a data entry plan (for paper-based surveys)

If the NO plans to use paper based tools (questionnaires) for data collection, the baseline tool should develop a data entry plan as part of the BDD. If data is collected using electronic tools, the baseline team does not need to prepare a data entry plan. Instead the team should agree on key steps they will take to ensure collected data is efficiently consolidated, cleaned and analysed.

For a paper-based survey, the data entry plan will outline the steps and processes the baseline team will follow to enter data. This plan should be developed during the baseline planning phase to enable baseline teams to enter data during the baseline or immediately after data has been collected. Electronic data entry formats should be developed for both qualitative and quantitative data using software packages with data entry builder modules or those specifically developed for data entry. Unless the data collection tool is in a tabular format, spreadsheet type entry designs such as Excel are not recommended especially for tools like the caregiver survey. The use of software such as CSPro and EpiInfo and other software packages which possess data entry capabilities, is highly recommended. SPSS has a data entry builder module but it is very expensive. Proper planning for data entry is very important to ensure data quality. When planning for quantitative data entry, the baseline team should consider the following tips.

- Select the most appropriate IT platform for data entry based on convenience of data entry, cost and availability of technical support to design data entry templates.

- Agree on key variables that will be used to code each key questions, e.g. 'ageU5' can be used for 'age of child under 5 years.'
- Agree on which baseline team members will enter data. In case the baseline team plans to have more than one person entering data, the team should agree on key steps that will be taken to check data for consistency and merge the dataset at analysis stage.
- Build quality checks into the data entry template. For example, the template can be formatted to only allow expected values to be entered. For example, if with a gender question it is agreed that 1= male and 2 = female; the data entry template can be formatted to only allow 1 or 2 to be entered for this variable.

A good qualitative data entry plan should:

- Use entry codes that match the questions that will be used (as written for focus group discussion (FGD) and key informant interviews (KII)).
- Contain a matrix that can be used to categorise key responses by theme (or group of respondents – boys or girls).
- Enable baseline teams to add or amend themes that emerge after data collection.
- Include one matrix per community or cluster, using the same template – this will enable baseline teams to trace back key issues to a certain community or gender/group.
- Include a summary template that can be used to aggregate data from all communities/respondents organised around the key themes that emerge from data collection.

## Phase 1 Step 11: Pre-test baseline design

It is very important to pre-test all components and steps of the baseline process because it is costly, time-consuming, and in some instances, not even possible to correct mistakes once the full baseline study is launched. Ideally, the baseline pre-test should be structured and organized as a learning lab, to enable key baseline stakeholders to experiment with all components and steps of the baseline process which include: community mobilisation, setup of data gathering tool (ODK, paper questionnaire etc.), enumerator and encoder training, data gathering, quality control, data entry and cleaning and data analysis and validation.

The baseline team should ensure that baseline pre-testing is planned well to enable all stakeholders to participate in the pre-testing and in the primary analysis of findings and use of findings, in order to make necessary adjustments to the sampling methodology and tools. The following tips should be applied.

- Before pre-testing the baseline tool, ensure that the full baseline methodology is agreed upon with stakeholders and all key tools developed.
- When pre-testing the data collection tools (both qualitative and quantitative) and sampling methodology, select a smaller sample (e.g., 5 percent of APs) and implement the entire baseline methodology as outlined in the BDD.
- When selecting APs for pre-testing, it is recommended to include APs from different contexts. In each of the pretesting APs, smaller samples (approximately 20 persons/other study units e.g. schools) should be sufficient to help the baseline team identify strengths and gaps in the baseline design and make the necessary adjustments to those designs. Where possible, the baseline design should be pre-tested in the communities where the actual baseline data will be collected.

The checklist below contains key questions which can guide the baseline team to effectively plan for pre-testing a baseline design.

**Key questions during pre-testing of baseline methodology and tools**

| Identified problems | Actions for improvement |
|---|---|
| **Sampling** | |
| What component of the sampling process did we struggle with? | What adjustments to the scoping process should be made before actual data collection starts? |
| What gaps did the sampling methodology have (key issues not thought about/included in the baseline design? | How will these gaps be addressed? |
| ***Data collection tools*** | |
| Which parts of the questionnaires / and qualitative interview guides had problems? | How will these problems be addressed? |
| Which questions did interviewers struggle to understand/ask? | What adjustments should be made to these questions? |
| Which questions did respondents struggle to understand / answer? | How will these problems be addressed? |
| Which questions were sensitive to respondents? | What adjustments should be made to these questions? |
| Which aspects did we struggle to observe? | What adjustments need to be made? |
| Which sections of interviews turned out to be very long and boring? | What is the best way to adjust or energise these sections? |
| ***Equipment and Logistics*** | |
| Which components of our logistics plan (transportation, housing, etc.) did we struggle with? | What changes should we make to improve the logistics plan? |
| Which equipment did we struggle with? Which additional equipment had we not thought about? | How will challenges related to equipment be addressed before actual data collection commences? |
| How did the mobile data collection devices perform? What challenges, if any, did we face with these devices? | What adjustments do we need to make to mobile data collection devices before actual baseline data collection commences? |
| ***Personnel*** | |
| Which team members struggled to understand and execute their roles? | What adjustments need to be made on the team/roles/facilitation to enhance team effectiveness? |
| Which team roles were not facilitated well? | |
| Which additional roles had we not thought about before? | |

Based on the results of the pre-test, the baseline team meet to finalise the baseline design document. The focus of revisions, if needed, would include but not be limited to the following:

[1] baseline purpose and objectives,

[2] baseline scope

[3] data gathering techniques

[4] data gathering instruments

[5] data analysis plan

[6] data quality control plan

7] data validation plan

[8] baseline sites

[9] schedule

[10] budget.

Once completed, the baseline team seeks approval from the National Director, and then shares the approved BDD with the RO, SOs and GC.

# PHASE 2: BASELINE EXECUTION PHASE

## An approved baseline design document guides the baseline team during the process of implementing the baseline. During this phase, the baseline team should follow these five key steps.

### Phase 2 Step 1: Organise baseline execution team

Once the BDD has been endorsed by the PST and GC, and approved by the national director, the first step in this phase is to organise the baseline execution team. It could have the same composition as the baseline planning team (NO DME, sponsorship and grants staff, RO, SO and GC), with the addition of members from the APs/clusters (such as AP supervisors, community leaders, primary partners like government entities, health workers and teachers). It is important to clearly define the role of each team member and ensure their availability for the entire baseline execution phase.

### Phase 2 Step 2: Conduct sampling

After the baseline team has identified the sample size to be selected from each AP, the next important task is to determine how samples will be selected for the baseline survey. As indicated earlier, there are a number of tools used for data collection during baseline survey. Here is the list of the tools with recommended sampling methods.

| Tool | Sampling Method |
|---|---|
| Caregiver Survey Questionnaire (CGS) | Two-stage cluster sampling is most commonly used. But where possible, other sampling methods can be used. |
| Youth Healthy Behavior Survey (YHBS) | The recommendation is to combine YHBS survey with the caregiver survey so that youth are selected from the households randomly selected for the caregiver survey |
| School-based Test About Reading (STAR) | Stratified random sampling (schools are considered as strata). In some cases cluster sampling may be considered. Refer to STAR sampling guide (Annex 5). |
| Development Asset Profile (DAP) | Similar to Youth Healthy Behavior Survey |
| Anthropometric survey (can be combined with caregiver survey) | Two stage cluster sampling (combined with caregiver survey) |
| Qualitative methods such as FGD or KII | Non- probability sampling |

Youth Healthy Behavior Survey (YHBS), Development Asset Profile (DAP) and anthropometric surveys can be combined with the caregiver survey as far as sampling methods are concerned. This means youths (for YHBS and DAP) and children under 5 years of age (for anthropometric survey) can be taken from the same households selected for the caregiver survey. For School-based Test About Reading (STAR) and other school-based tools, refer to their respective guidelines (Annex 5 for STAR tool) for details in sampling. LEAP 3 encourages baseline teams to consider the following tips during the application of a 2-stage cluster sampling approach.

1. Determine the total number of clusters and number of households per AP based on the sample size to be collected per AP. For example, if the sample size per AP is 300 households, then data from 300 HHs can be collected using: (a) 20 clusters x 15 households OR; (b) 30 clusters x 10 households. It should be noted that a cluster size (number of households per cluster) should not be greater than the total number of clusters.
2. Compile sampling frames for each AP selected to participate in the TP baseline. Since, in most cases, it is difficult to have a complete list of all households in an AP, the sampling frame will be the list of primary sampling units (the smallest administrative units such as villages where the AP plans to implement the TPs) and the number of households in each primary sampling unit. Where such lists do not exist (or are out of date), baseline teams are encouraged to work with government officials and community leaders to develop a list. The table below (example 6) shows the basic format of a sampling frame.
3. The first column of the sampling frame contains serial numbers which start from 1 and continue to the end of the list. The second and third columns are obtained from the list prepared in step 2 above. The third column (cumulative # of HHs) enables baseline teams to calculate the cumulative number of households and write these in each row. The cumulative number of HHs is computed by adding the number of households (in that row) to the sum in the previous row, beginning at the top and working down to the last one.
4. Calculate a sampling interval (SI) by dividing the total number of households by the total number of desired clusters from the AP. When there is a sampling interval with decimals, don't round up. Just use the whole number.
5. Randomly select a start number (SN) between 1 and the sampling interval (SI). Put that number in the fifth (calculation) column in the row corresponding to the minimum cumulative number greater or equal to the randomly selected number.
6. Determine subsequent clusters by adding the sampling interval (SI) to the start number which was randomly selected in the step above. Enter the sum in the fifth (calculation) column in the row corresponding to the minimum cumulative number greater or equal to this sum. In some cases, the same row may be repeated two or more times. The number of times a row is repeated shows the number of clusters which must be selected from the respective primary sampling unit (village).
7. Continue in a similar manner to identify the remaining clusters by: adding the sampling interval to the previous result; finding the minimum cumulative number greater than or equal to that number and entering the number in the fifth (calculation) column in the row corresponding to that cumulative number. Continue until the required number of clusters has been assigned.
8. Once step 7 is completed, count the number of clusters required for each primary sampling unit and write the numbers in column six (# of clusters). Note that larger sampling units have more clusters. This is because this process uses a technique called 'Probability Proportional to Size (PPS)' which ensures that more clusters and households are selected from sampling units which have high populations (than those with smaller populations)

*Example 6: Sampling frame format*

| # | Lowest Admin unit (village) | # of HHs | Cumulative # of HHs | Calculation | # Clusters |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | | | | | |
| | Total | | | | |

**Selecting households in a cluster**

After the baseline team has identified the primary sampling units from which clusters will be selected, the next step is to outline the process of selecting households (secondary sampling units) in the selected clusters. Ideally, households that will participate in the survey should be randomly selected from a complete list of all households in the villages or towns selected. However, in most places where WV works, random selection of households is not always possible. To address this challenge, the baseline team should apply the following approaches.

- The first (and preferred) approach is called 'segmentation'. Segmentation allows baseline teams to randomly identify a segment (or cluster) of houses in the primary sampling unit. This is done by dividing the entire primary sampling unit (village) into roughly equal segments containing the number of houses that needs to be surveyed. One segment is chosen at random and the required number of houses (per cluster) will be surveyed.

- The second method for selecting households within a cluster is the random walk method, which is used in EPI cluster surveys. It can be used as an alternative to the segmentation method when the village or town cannot be segmented accurately, or perhaps when it has not been possible for the baseline team to generate a map of the survey area. This is done during actual data collection by going to the centre of the cluster (or an important location such as a school, market or institution) and randomly selecting a direction in which to proceed. Random direction is found by spinning a pen or pencil in the air and letting it land on the ground, pointing in the direction to proceed. Baseline teams then interview respondents from the first household and continue in that direction either until the team reaches the border of the cluster or until they get the desired number of households – whichever comes first. If the number of households in that direction is less than required, the baseline team goes back to the central point and spins a pen to choose another direction and then continues the survey in the new direction until they get the desired sample size per cluster.

**Tips – lessons learned from LEAP 3 baselines indicate that:**

- It is easier to conduct a two-stage clustering approach, by first sampling from the total clusters and then from the HHs within the cluster. It is better to have more clusters with fewer households/units. In any case the number of households per cluster should not be greater than the total number of clusters in an AP.
- The number of HHs per cluster should be the same in each cluster.
- It is recommended that a cluster should have between 70-120 HHs/Units to select from.
- WV uses the DHS general recommendation of a sample size of 20-30 HHs or Primary Sampling Units within the cluster

## Phase 2 Step 3: Hire and train enumerators and data handlers

Effective data collection for programme baselines requires the coordinated efforts of many players. Vital to this process are the enumerators who interact with and collect baseline data from programme beneficiaries. Baseline coordinators should ensure that the required enumerators (both in number and their competencies) are carefully recruited and effectively trained. WV recommends that enumerators are only recruited after the baseline design has been completed and approved, and draft data collection tools have been developed. Recruiting enumerators at this stage enables programme teams to ensure recruited enumerators receive adequate training/orientation on all core components of the baseline design and methodology including: (a) baseline purpose/objectives; (b) sampling methodology; (c) data collection tools and protocols; (c) data quality assurance/spot checks; (d) data entry and cleaning processes; (e) problem solving skills – including handling of missing cases/refusals/complaints; (f) child protection and other WV mandatory standards. Enumerators should be actively involved in the pre-testing of baseline design, methodology and tools. This helps them build the skills and competencies necessary to enhance their effectiveness during baseline data-collection.

Decisions regarding recruitment of enumerators should be made by the baseline team based on:

(a) technical skills/experience required

(b) language skills

(c) time needed to complete the baseline task

(d) available resources to compensate them.

WV encourages its staff and partners to consider using students, interns or volunteers as enumerators. In addition to helping students/interns build research skills, this strategy could help reduce the costs of baselines.

The baseline team works with the enumerators to ensure they are fully competent to administer the data collection tools, including what to do when a respondent is not available or refuses to participate.

Please see Annex 3: Training outline (for enumerators training).

## Phase 2 Step 4: Prepare for fieldwork

The baseline execution team should make sure that all transportation, housing and other arrangements are made before the actual baseline dates. The necessary equipment should be acquired and pre-tested before actual data collection. This includes laptops, tablets, printers, back-up power sources, mobile phones, writing and recording implements and consent forms etc. In addition, all partners and communities who will participate in the baseline survey, should be mobilised before actual data collection starts.

## Phase 2 Step 5: Gather qualitative and quantitative data

**Daily field deployment:** Before enumerators, FGD facilitators and interviewers are deployed in the field, supervisors need to ensure that each of them has the list of respondents, data gathering tool(s), consent forms, letter of introduction (if needed), means of identification, means of communication, drop-off and pick-up times and basic provisions.

While data gathering is taking place, supervisors should monitor the times and any communications from the enumerators. Transportation should be on standby, in case there is a need for supervisors to follow up, or cover for enumerators and/or to address early or late completion of data gathering tasks.

**Informed consent for respondents:** The principle of respect for persons incorporates two fundamental ethical considerations: [a] Respect for autonomy, and [b] protection of vulnerable persons. These are addressed by individual informed consent procedures that ensure that respondents understand the purpose of the measurement and that their participation is voluntary.

## Phase 2 Step 6: Conduct quality control

Supervisors should check the number of completed questionnaires, interviews and discussions at the end of each day. Any early completions or delays should be addressed (e.g., why, by whom, how often) and schedules adjusted accordingly. The contents of completed questionnaires, interviews and discussions should also be checked at the end of each day to check for completeness and legibility. Quality control measures embedded in the data entry templates should also be monitored.

Any quality-related issues should be addressed as soon as possible, and before leaving the field site. Please see Annex 4: Data collection management tools.

## Phase 2 Step 7: Enter and clean data

During programme baselines and evaluations, errors often happen regardless steps taken by the study team to efficiently plan, collect and manage data. Errors can be introduced at any stage of baseline planning, data collection, data entry and storage. The use of modern electronic devices is also still susceptible to errors. As with any other experiment, when errors happen and are not detected and corrected, those errors compromise the quality of research findings and conclusions. Therefore, it is essential that both quantitative and qualitative data is screened for errors (and those errors corrected) before data analysis begins. Data cleaning refers to efforts taken by the researcher/team to screen collected data, identify and fix any errors before the data set is analysed. Below are some practical steps that can be taken to detect and fix errors in baseline/evaluation data.

**Quantitative data cleaning:** Cleaning quantitative data is straightforward. It requires the research team to pay attention to the following key areas.

- **Missing information:** During the course of data collection and entry, baseline teams are encouraged to continuously check their questionnaires and data entry files to identify and fix any missing information. Missing information could include: questions that were not adequately answered; variables in the data entry set that are missing (especially where the task of data entry is divided/handled by different people); and/or missing values in both manual and electronic datasets. There are many reasons that can lead to information missing in a dataset. The most common include: loss in a sample (targeted interviewees could not be reached or declined to participate and were not replaced); poor recall by respondents (a farmer interviewed could not remember the quantity of cereals his family harvested); and equipment failure (weighing scale failed during the process of weighing children U5). Every effort should be taken by the research team to correct missing information before data analysis begins. This could be done manually or electronically. DME specialists are encouraged to explore advanced statistical techniques (i.e. creation of dummy variables, transposition of variables, etc.) that can be applied to enhance data cleaning rigour.
- **Identify outliers:** In statistics, outliers refer to values that seem inconsistent with the rest of the values recorded for a given variable. To identify outliers, the research team can quickly work out 'minimum' and 'maximum' values and establish a predictable range of values for key variables in the dataset._Here is an example to illustrate this point. Suppose an AP in Mozambique decides to collect anthropometric data from 400 children aged 6-12 months. The AP team quickly checks the data and determines that the normal weight for more than 90 percent of children studied ranged between 6.5kg and 13kg. Any children in the sample with recorded weights way out of this range (e.g. 2kg or 27kgs) would then be considered outliers. Such cases could have been recorded

wrongly or could represent cases of extreme but real biological variation (such as acute malnutrition or severe obesity). These cases should either be fixed (errors corrected) or excluded from the analysis – to avoid skewing (inflating or deflating) key measures of central tendency i.e. the sample mean.

- ***Checking for inconsistencies in the data:*** Sometimes inconsistent information can be observed in the data. For example, it may be reported that a child under 5 years of age has completed primary school. Or that the age of a child is greater than their parent…etc. Cross-tabulation of variables can help in identifying such inconsistencies.
- ***Checking for out-of-range values:*** Another type of errors committed during data collection or data entry is the capturing of out-of-range values. For example if the code for gender is given as female = 1 and male = 2, these should be the only values seen in the data for gender. Generating frequencies of the variable (gender in this case) can help discover out-of-range values.

**Qualitative data cleaning:** Cleaning of qualitative data should be guided by the same principle - to identify and correct any errors in collected data. As with quantitative data, errors can be introduced at any stage of planning, collection and management of qualitative data. Because qualitative data is often unstructured, LEAP 3 recommends that baseline teams should progressively clean qualitative data as it gets collected. There are many techniques that can be used to clean qualitative data. For example, the baseline team could develop a framework to summarise qualitative data collected on a daily basis. Daily team meetings are encouraged to enable baseline teams to check their records of data that has been collected/documented; to compare notes taken by different team members and to summarise collected data into emerging themes, trends and any interesting quotations.

# PHASE 3: DATA ANALYSIS AND UTILISATION PHASE

## Phase 3 Step 1: Organise the data analysis and report writing team

LEAP 3 recommends that a national office should designate a team to manage baseline data analysis. This team should include members of the baseline team (who participated in the actual collection of data) plus any additional people with excellent data analysis skills.

## Phase 3 Step 2: Analyse the data

Data analysis refers to the application of different statistical or non-statistical techniques. It follows a set process to make use of the collected data and use the results to support decision-making. Data analysis is guided mainly by the objectives of the study. In TP baselines, the minimum objectives are to:

a. Set benchmarks against which programme impact can be monitored and evaluated.
b. Validate and strengthen programme targeting and technical approaches. Triangulation and other data validation approaches may be required to arrive at appropriate conclusions during data analysis.
c. Strengthen understanding of existing needs/opportunities and draw appropriate conclusions from the data analysis. This may require analysing relationships and dependencies between relevant variables.

The following steps are recommended for baseline data analysis:

**Descriptive analysis:** The first step that baseline teams should take is to develop descriptive information to summarise the important estimates of key variables. Descriptive information should be generated for both qualitative and quantitative data.

For quantitative analysis, a helpful starting point would be for baseline teams to create dummy tables that outline the different variables and analysis required for each variable. Creating dummy variables before analysis begins, will help baseline teams identify key questions to prioritise during descriptive and detailed analysis[11]. There are two types of descriptive statistics that should be generated for all quantitative data collected to estimate LEAP 3 baselines. These include:

a. **Frequency Distributions:** In statistics, frequency distributions describe the rate at which different values associated with an important variable occur in the sample. Frequency distributions are most commonly computed for categorical data (variables that take on known values such as gender and 'yes or no' variables). Every statistical software package provides practical guidance that should be followed by baseline teams to compute these frequencies. Frequency distributions and answer basic analyses questions such as what proportion of respondents were male/female, the completed different levels of education, access to safe water and sanitation, and so on. This type of information helps baseline teams to quickly understand the basic characteristics of target populations, coverage of key services and gaps. Frequency distributions can either be documented as figures or charts.

b. **Measures of central tendency**: A measure of central tendency is a single value that can be used to describe a set of data by identifying the central position within that set of data. The most common/useful measures of central tendency for community development baselines could be arithmetic mean, mode and median. All these measures can be easily generated using basic statistical software (which provides adequate guidance regarding the steps that need to be taken by

[11] The internet contains many free references on creation of dummy variable and dummy tables which baseline teams should consult for assistance on this subject.

baseline teams wishing to compute measures of central tendency). The table below explains each measure of central tendency.

| Measure of central tendency | Characteristics |
|---|---|
| **Mean** | • The mean of a set of values is the sum of the values divided by the number of the values.<br>• Mainly used for continuous data but can also be applicable for discrete data.<br>• It is influenced by outliers (extreme values) in the data.<br>• Preferred use is when the distribution of the data is not skewed. |
| **Median** | • The median is the middle value in a set of values arranged in order of magnitude.<br>• The median is not influenced by outliers.<br>• Applicable for both discrete and continues type of data.<br>• Mostly preferred when the distribution of the data is skewed. |
| **Mode** | • The mode is the most frequent value in a set of values.<br>• If values are displayed on a chart, the mode represents the highest bar in a bar chart.<br>• Mainly used for nominal data. |

Descriptive information can be generated by organising transcribed data (either manually or with software), coding data and using an appropriate framework to organise the data into recurring themes, emerging patterns and key quotes. Enormous literature exists on this subject (and should be explored by DME specialists who lead baselines in national offices).

**Second-level statistical analyses:** Conducting descriptive analyses provides useful but not necessarily sufficient information to assist programme staff and partners in gaining an in-depth understanding of target communities. Such analyses will help validate/adjust proposed technical approaches and guide investments towards interventions that will maximize the well-being of children. Descriptive statistics can only estimate the current values of key indicators but do not help programme teams to adequately analyse the relationships between key indicators/variables. This inference requires a second level of statistical analysis which can be used to test the validity of research conclusions. There are many statistical computations that can be done to test the existence and direction of relationships between key variables (indicators). The following are some of the recommended methods of analysis:

a. Cross-tabulation: In statistics, cross-tabulation of variables is used to determine the possible existence of a relationship between key indicators or variables representing those indicators.[12] Also known as 'contingency table analysis', cross-tabulation uses frequency distributions to show the relationship between different variables in the dataset. Cross-tabs can easily be generated by most statistical software and can be used to compare one variable against another. For example 'farmers adding value to their crops' and 'profitability from those crops'. Cross-tabs can also be generated to show a relationship between one variable (e.g. prevalence of diarrhoea) and multiple variables (e.g. access to safe water, hygiene and sanitation behavior or education level of mothers/caregivers). However, the results of cross-tabulation analysis need to be verified using measures which test the statistical significance of such identified relationships. The most commonly used test to assess the statistical significance of cross-tabulations is the Chi-square. The Chi-square tests whether or not two variables are independent. If they are independent (or have no relationship between them), the results of the chi-square statistical test will be 'not-significant'. If the variables are related, chi-square results will be 'statistically significant'.

[12] Cross tabulations are normally computed for categorical variables. Computing cross tabulations for continuous variables requires grouping of those variables (all major statistical software provide guidance on grouping of variables).

b. Computing correlation coefficients is another technique that baseline teams can use to determine the existence and strength of a relationship between variables/indicators of interest. For example, a baseline team might have collected data on key indicators such as the prevalence of diarrhoea among children U5, access to safe water by families where these children live and hygiene and sanitation behaviors practiced by mothers/caregivers of children U5. By computing correlations between these three variables, a baseline team can easily establish whether or not a relationship exists between the prevalence of diarrhea and access to safe-water/sanitation.

Most statistical software such as SPSS has inbuilt capability to test the statistical significance of the existence of correlation.

### The principle of weighting

During sampling, the population sizes of different groups should be considered, in order to avoid the sample over- or under-representating certain groups in the population. For example the same sample size might be taken in all APs and the data then aggregated to generate TP level values. If the population sizes among the respective APs varies, then the TP level baseline value may not be representative of the population. Applying weights to samples can help address such issues to improve the representativeness of the sample.

**Types of weighting:** There are many types of weighting that can be used to minimize sampling errors that could be inflated due to under- or over-representation of different population groups. Two types are applicable during TP baselines, - self-weighting and design weighting.

**Self-weighting:** If the TP level sample size is proportionally allocated to APs during sampling, there will be no need of weighting during analysis as the sampling process allows self-weighing.

**Design weights:** In cases where sampling is undertaken without taking the population proportion into consideration, sample data should be weighted using the inverse of the selection probabilities. This means: design weight = 1/selection probability. For example, if 150 households are selected from a total of 2500 households in AP1, and the same number of (150) households from a total of 3500 in AP2, then the weights will be 2500/150=16.67 for AP1 and 3500/150 =23.33 for AP2. Demonstration of the application of design weights is shown in Annex 6.

## Phase 2 Step 3: Validate the baseline results

In research, validation refers to a process undertaken by a research team and stakeholders to establish the soundness, accuracy or legitimacy of research findings. LEAP 3 recommends that baseline results should be validated by all key stakeholders at all levels of the programme (i.e. local and national levels). In line with DPA principles, baseline teams should ensure that community level partners (including children) are actively involved in the validation of baseline results. At local and national levels, validation of baseline results helps WV staff to establish the credibility of the baseline findings and engage key stakeholders in interpreting the baseline results, including drawing conclusions and generating recommendations.

Meaningful validation of baseline results requires prudent planning. LEAP 3 recommends that after data analysis is completed, baseline teams should carefully reflect on their key stakeholders and use the preliminary baseline findings to develop communication products that will engage all stakeholders in discussion and validation of baseline results. Communication products could include summary reports, baseline fact sheets, oral presentations or posters. Whatever communication product is chosen by the baseline team, it is important for all stakeholders to receive a briefing on the baseline: (a) purpose/objectives/questions; (b) methodology and; (c) key findings – both qualitative and quantitative.

During the validation process, the baseline team should engage key stakeholders to reflect deeply on key baseline findings focusing on:

- **Fostering a clear understanding** of baseline findings (including what stands out, what is interesting, what doesn't make sense, what stakeholders disagree with and what requires further investigation).
- **Generating consensus** on proper interpretation of baseline results including: (a) the current state of key CWB indicators, (b) generating factors that contribute to the current state of key CWB indicators, (c) agreeing on key differences in vulnerabilities (or strengths) that exist across different target populations and across different Aps and (d) assessing key external factors that influence the current state of key baseline findings.
- **Triangulation of baseline results.** In social research the accuracy and reliability of the results of any study must be validated using the results of similar studies conducted on that subject. This process is called triangulation. Triangulation helps baseline stakeholders to identify baseline findings that can be confirmed and those which may require further discussion or investigation.
- **Generating key conclusions and recommendations.** This includes answering the following questions:
    a) What are the most crucial development challenges highlighted by the baseline results?
    b) What implications do the baseline findings have for the root-cause analysis that was conducted for programmes that have been baselined?
    c) What priorities/project models need to be emphasised by the different TPs baselined?
    d) What crucial partnerships should be strengthened by WV in order to enhance delivery and impact of TPs?
    e) What implications do the baseline findings have for programme monitoring and evaluation?

## Phase 2 Step 4: Write the baseline report

**Process: The baseline report is crafted by the national office and reviewed/endorsed by the PST and Global Centre E&L.**

A baseline report is used to communicate baseline findings. The purpose of this report is to present the main findings (including conclusions and recommendations) from a baseline survey. The key audience for the report include: community-based partners (including children and organisations WV partners with), WV audiences (NO leadership/technical staff, regional office and GC technical specialists, support office technical and programme managers and government leaders). A good baseline report should include the following sections:

A. **Acknowledgement** – listing and giving credit to all key stakeholders who participated in baseline planning, data collection, analysis/validation, etc.

B. **Executive summary** – (also known as a management summary) this component of the baseline report presents the key highlights of the baseline survey (both process and key findings). It is normally prepared for an audience of key decision makers who need to appreciate the baseline process and understand key baseline findings, but don't have time to plough through the entire report. A good executive summary should include: (a) a brief background description – including baseline purpose/objectives/questions, (b) summary of the methodology and (c) highlights of key baseline findings, conclusions and recommendations.

C. **Table of contents** – which outlines all the key sections/chapters contained in a report and helps a reader looking for specific information, to quickly locate the appropriate section where that information is contained.

D. **List of acronyms** – spelling out all the key abbreviations used in the report. Please note in professional writing, abbreviations should only be used after they have been defined. For example, before using AP in a baseline report, the writer should use 'Area Programme' first and then introduce AP in brackets.

E. Chapter 1: **Background and context:** this should summarise key development issues in the country, NO strategic priorities, description of approved TPs that have been baselined, etc.

F. Chapter 2: **Baseline purpose and objectives** – this should outline the baseline purpose, objectives, and questions (including research questions).

G. Chapter 3: **Methodology:** this should highlight: (a) the study design' (b) types of qualitative and quantitative data collected, (c) techniques employed to collect that data, (d) data sources and target populations, (e) sampling techniques, (f) sample size and how it was determined, (g) data collection tools used, (h) data quality assurance and analysis approaches that were used etc; (i) any study limitations, sources of error, and areas that need additional investigation. The methodology section should also include key research questions and hypotheses investigated during the baseline process.

H. Chapter 4: **Key findings:** this section should be organised according to the key research objectives and questions. It should include descriptive results that show the current status of key CWB indicators and other indicators (e.g. community resources, structures, and sustainability, faith and development issues) that were assessed by the baseline survey. Descriptive statistics should be supported with qualitative findings (key themes, quotes). Descriptive statistics should be presented using tables, figures and charts that help reviewers of the baseline report to quickly understand the numbers and the conclusions behind those numbers. Whenever used, tables and figures should have legends explaining what they are summarising.

I. Chapter 5: **Discussion and conclusions:** In this section the baseline team should present an interpretation of the results and major conclusions. The discussion section should compare the baseline survey results with the results of other studies (using the validation and triangulation exercise conducted by key stakeholders).

J. Chapter 6: **Recommendations:** key recommendations are drawn from the baseline findings to outline key steps programme staff and partners would need to take in order to enhance the overall programme effectiveness. They should validate and confirm the components of the programme strategy (including root cause analysis, proposed interventions, etc.) which are relevant and outline gaps in programme design, implementation and M&E plans that need to be addressed.

K. **References:** all technical resources (published and unpublished) consulted during the process of planning and conducting the baseline survey should be listed. Failure to list and give credit to key references could infringe on copyright which can have legal implications. For all key references, baseline teams are encouraged to state: the author's name (last, then first), title of the book or paper, edition, place of publication, year of publication.

L. **Appendices:** important documents/resources used (and/or developed) during baseline planning and implementation should be annexed to the report. Typically, these would include: copies of key data collection tools, terms of reference for consultants (where applicable), the profiles of key baseline team members and copies of tables and figures generated during analysis.

## Phase 2 Step 5: Use baseline information

After the baseline report has been finalised and approved by the ND, the programmes director should lead a process of using the baseline results to set performance targets for all outcome indicators and make any necessary adjustments to TP designs. National offices are encouraged to refer to LEAP 3 guidance on setting performance targets.

# ANNEXES

## ANNEX 1: Sampling guide for special groups or direct beneficiaries

### Applicable for sampling of RCs or MVCs

This purpose of this sampling guide is to outline the steps required to take a sample from a particular group of interest that is believed to benefit from project/programme interventions. This is an independent sampling procedure and the sample cannot be combined with other samples collected through other tools such as caregiver survey.

**Factors to consider when determining sample:** When a sample size is calculated, it is important to take the following into consideration.

1. **Number of special group** — What is the size of the target population (for example, the number of adolescents targeted)?
2. **Margin of error** (Confidence Interval) —The confidence interval determines how much higher or lower than the population average the sample average will be allowed to fall.
3. **Confidence Level** —The most commonly considered level of confidence of the actual mean falling within the confidence interval is 95%.
4. **Standard of deviation (proportion)** — How much variance is expected in the responses? If there are no estimates of the variance, using 0.5 will give the highest sample size.
5. **Sampling method** – which sampling method will be used? Such as simple random sampling or other sampling methods such as stratified, cluster…etc.

**Formula to calculate sample size:** To calculate the sample size based on the sample required to estimate a proportion with a certain confidence level, and assuming simple random sampling without replacement, the following formula can be used:

$$n = \frac{z^2 pq}{d^2}$$

Where

Z = value of standard normal variate for desired level of confidence (Z=1.96 for 95% confidence level)
*n* = necessary sample size,
*p* = proportion of the population having the characteristic,
*q* = 1- *p* and
*d* = the degree of precision (margin of error).

The proportion of the population (*p*) may be known from previous surveys or other sources. If it is unknown use *p* = 0.5 which assumes maximum heterogeneity (i.e. a 50/50 split). The degree of precision (*d*) is the margin of error that is acceptable by the survey team. In many studies *d* = 0.05, which would give a margin of error of plus or minus 5%. As the value of d is reduced, n increases.

**Sample size for finite population:** The above formula (1) for the sample size necessary for estimating a population proportion p, works well when the population in question is very large and the sample size is less than 5% of the population. If there are smaller, finite populations, and the sample size is 5% or above of the population, then finite population correction factor (fpc) is introduced to adjust the sample size.

The following formula (2) is used to adjust the sample size if the initially calculated sample size using the above formula is 5% or above of the population.

$$n_a = \frac{n}{1 + \frac{n-1}{N}}$$

Where:

*n* = initially calculated sample size as per the above formula (1)
$n_a$ = the adjusted sample size
*N = the size of the target population*

**Proposed options of sampling for adolescents survey:** If the population size of adolescents who are the target of a (TP) is 400 per (AP), here are the options of sampling:

**Option 1:** 400 is generally felt to be a small population and as long as each AP is expected to handle the survey, it can be considered feasible to include all 400 in the survey. This will depend on whether adequate enumerators can be hired and trained. So the first option is to include all adolescents in the survey.

**Option 2:** If there are logistical challenges such as the challenge of getting adequate number of enumerators in the AP area or issues related to budget or other technical challenges, then the option is to go for sampling. It is for this option that the above formulas are considered. In order to calculate the required sample size using formula (1), required parameters must be set.

For 95% level of confidence, the value of Z is 1.96. If d=0.05 and it is assumed P is unknown and set as P=0.5. The sampling method is Simple Random Sampling without replacement.

Calculating for n then gives:

$$n = \frac{1.96^2(0.5)(0.5)}{0.05^2}$$
$$n = 384.16 = 385$$

The sample size 385 is more than 5% (96% to be exact) of 400. Therefore, we apply fpc to adjust the sample size. Using formula (2) above, we calculate the adjusted sample size as follows:

$$n_a = \frac{n}{1 + \frac{n-1}{N}}$$
$$n_a = \frac{385}{1 + \frac{385-1}{400}}$$
$$n_a = 196.4 = 197$$

Considering 10% non-response, we add 10% of 197 which is 19.7 = 20 to this adjusted sample size and we get 217 or rounding up giving us 220. So with the parameters given above, a sample size of 220 adolescents per AP can be sampled. If any of the above parameters (especially d and p) are changed, the required sample size will be changed.

**Stratification:** Stratified sampling procedure can be applied to make sure different groups are well represented. The easier procedure is to first categorize the target population according to the variable of stratification. Then proportionally allocate the sample size to each category. For example, if the variable of stratification is gender, and if there are 35% female and 65% male adolescents in the population, then 35% of the sample (77 out of 220) should be females. If the degree of vulnerability of MVC is considered as a variable of stratification, the same principle applies. It is important that the number of categories are as few as possible to avoid complications.

# ANNEX 2: Sample size calculator

Revised_Sample_Sizes_Calculator_June_

# Annex 3: Training outline (for enumerators' training)

## Introduction

1. **Ice-breaker for them to get to know each other and work as a team**
2. **Overview of the baseline process (presentation)**
    - Summary of basic information sheet
    - What tools will be covered in training
    - Roles and responsibilities
    - Importance of data quality
    - Who is responsible for data quality? I AM!
    - How to achieve data quality: follow protocols, respect, attitude of a learner, know the tools and your role.
3. **Communication skills and interview protocol**
    - Active listening, body language (activities)
    - Role of data collector.
4. **Importance of completed tools**
    - Each one is like a bar of gold – ensure each one is handed to supervisor
    - Supervisor returns all to lead supervisor - must be safely and securely stored!
5. **Child friendly approaches and ethics**
    - Child friendly approaches: ways to make children feel more comfortable
    - Child protection tips and signing of child protection forms
    - Sharing child friendly warm ups/games.

**For each tool**

1. Overview of tool (brief presentation).
2. Hand out tools, allow time to read.
3. Group or pair role play the tool – groups/pairs share feedback with person in 'data collector' role.
4. Facilitators observe / answer any questions .
5. Repeat until everyone has had a turn to be the data collector.
6. Have a few groups publically role play to whole group – get feedback from group on what they liked and also what could be done better.
7. Ask groups to generate:
    a. List of any questions or instructions to them that are incorrectly translated or confusing / unclear
    b. 'Top Tips' for how to use the tool for best quality data.

# Annex 4: Data collection management tools

## FIELD TEST – SUPERVISOR'S CHECK-LIST

**Quality check!** Randomly check completed tools to see if they are correctly completed and identify problem questions.

| Look for: | How to improve data collection: |
|---|---|
| Missing data – for example, questions missed or information not completed. | Change the layout of the survey to make it more obvious what to complete. Ask for ideas from data collectors about what might make it easier to complete.<br>Remind data collectors not to forget to complete EVERY relevant section. Ask them to come up with tips on how to ensure nothing is missed. |
| Data entered in wrong place. | Clarify how to complete the tool by showing real 'good' and 'bad' examples from the pile of papers, so people can see the difference. Ask them to come up with tips on how to ensure data goes in the right place. |
| Illegible responses. | Emphasise the importance of clear handwriting. Show people a good and bad example (for example scan some examples onto computer and project on a large screen. Ask people: which one is easier to read? Ask how handwriting can be improved? |
| 'Choose one' responses with more than one response given. | Put important instructions in bold like 'choose only one'.<br>Use one box to enter the single response. |
| 'Choose all that apply' questions with only one response given. | Put important instructions in bold like 'choose all that apply'.<br>List all possible options (for example, a, b, c, d in the answer box) so data collector can circle all that apply. |
| Skip patterns not followed – for example, a 'no' response meant skip next question, but question was asked and data entered. | Go over skip patterns again to ensure everyone is clear about what to do – test data collectors on the spot, by picking one out in the plenary and asking them: "if I say 'no' to this question, what do you do?" and so on. If they get it wrong, ask another data collector to give the right answer. Repeat until everyone is clear. |
| Questions commonly answered 'other'. | Get input from data collectors about what common responses were given instead. There are two options:<br>1. Agree on broader categorisation of responses using existing categories<br>2. Add more categories as options.<br>Example 1: If when 'other' is specified as 'work on family farm' - this could have been categorised as an existing option 'work for family business'.<br>Example 2: If when 'other' is specified as 'left the country' and this is a common response, consider adding this as a category. |
| Questions commonly answered 'don't know' or 'no response'. | Get input from data collectors on why this question is not meeting with a useful response. Is the wording confusing? Is it a controversial issue? There are two options:<br>Re-word the question to make it more appropriate<br>Remove the question as not valid. |

**Finally …**Update the translated tools the same day if possible and re-print sufficient copies, or update the survey online if using smart phones for data collection. Check back with the data collectors that they are ready and confident for the real data collection to start!

## Supervisors Field Guide

- ✓ Keep in contact with your enumerators throughout the day.
- ✓ Provide data collectors with sufficient copies of tools and pens (have spare ones ready).
- ✓ Ensure data collectors locate correct houses and children.
- ✓ Check through completed surveys as soon as you get them - data collectors can still return to the household and complete the survey correctly.
- ✓ Keep a list of completed records, HHs that refused or were not available.
- ✓ Check off each completed household or child on the list as you receive it.
- ✓ Observe the data collection process, especially with children, and note body language. Where necessary and without interrupting the process, give advice to the data collector to improve or address the situation.
- ✓ Be available at all times in case a problem arises.
- ✓ Keep a list of mobile numbers and names of all your enumerators.
- ✓ Remain in contact with the Team Lead and make appropriate decisions as issues arise.
- ✓ Once a cluster or group is finished, count and note the number of records carefully.
- ✓ Organise and store the records from one cluster in an envelope, wrap with an elastic band or use a spare sheet of paper to fold around them and keep them together.
- ✓ Collect all completed tools from data collectors and store safely.
- ✓ Seek feedback from the data collectors and keep them motivated!

| **Look for** | |
|---|---|
| ☐ | Missing data - questions missed/ not completed |
| ☐ | Data entered in wrong place |
| ☐ | Illegible responses |
| ☐ | 'Choose one' responses with more than one response given |
| ☐ | 'Choose all that apply' questions with only one response given – there should be a '0' or a '1' in every box |
| ☐ | Skip patterns not followed |
| ☐ | Questions commonly answered 'other' |
| ☐ | Questions commonly answered 'don't know' or 'no response' |

- ✓ Keep in contact with supervisors and listen to their feedback.
- ✓ Have a list of all supervisor's names and mobile phone numbers.
- ✓ Quality check on a daily basis – afterwards is too late.
- ✓ Ask what common errors or issues are coming up – is there something that can be done to address this?
- ✓ Work with supervisors and coordinator to make critical decisions as issues arise.
- ✓ Communicate with baseline coordinator on major issues and on a daily basis.
- ✓ Enumerators, Supervisors and YOU are all responsible for data quality
- ✓ Randomly check a sample of surveys from each supervisor's group and give feedback for improvements.
- ✓ Talk to data entry supervisors to understand common errors and communicate this immediately with all supervisors to check for this.
- ✓ Make sure supervisors communicate with all their enumerators clearly and urgently on any changes or issues to be addressed.
- ✓ Supervise your supervisors – talk to enumerators directly too.
- ✓ Keep supervisors motivated!

## Six common errors to check for:

| **Missing fields (missing data) including ADP and HH #** | **Writing not legible** | **Data entered in wrong place or skip not followed** | **Questions commonly answered 'other' or 'don't know'** |
|---|---|---|---|
| **Choose ONE questions with more than one answer given** | | **Choose ANY (many) with only one – there should be a '0' or '1' in every box** | |

## At the end of every day:

- Ask supervisors to:
  - count the records carefully and check for any missing against their list, before releasing enumerators.
  - Thank enumerators for their time and efforts in contributing to this important process.
- Hold a brief review to listen to feedback; discuss and decide together on any issues that come up.
- Make sure everyone is absolutely clear on any changes required.
- Work with supervisors to organise the records by HH #, cluster and ADP.
- Store the records safely together (e.g. use large envelopes or elastic bands, and store in boxes or folders) – do not leave them lying around.
- Keep a master list of # records completed in each cluster and ADP, by tool (CGS & STAR) and share with baseline coordinator.
- Remember completed records are like bars of gold!

# Annex 5: School-based sampling for literacy assessment

**EdLS Goal: To contribute toward achieving improved learning outcomes for all children in areas where World Vision works.**

**This sampling design can be used for all school-based assessments (literacy or numeracy) regardless of the tool being used (STAR, EGRA, LB or MoE validated tool).**

### Introduction

Under LEAP 3 programming, integrated baseline and follow-up evaluations will be conducted at national level using a modular approach, to allow for the inclusion of all technical programmes (TP) being implemented in each Area Programme (AP). This is intended to facilitate data collection and analysis for key indicators on Child Well-Being (CWB) Objectives. Sampling will therefore be done at AP level. Every AP with a technical programme is eligible for baseline and evaluation. Depending on the sample size, this will provide data that can enable change analysis at AP level, cluster/regional and national/TP level. It also allows for standardised reporting on CWB objectives.

### Purpose

This guide provides a simple sampling approach to help national offices select an optimum sample size for measuring literacy outcomes for children enrolled in and attending WV supported schools. The resulting sample should allow sufficient assessment data to be collected to estimate or detect changes in learning outcomes and determine the effectiveness of education programming. The data should also provide adequate information for decision-making on education programmes.

The articulated sample design is for AP baseline and subsequent evaluations related to the education child well-being objective:

***"Increase in primary school children who can read"***

| Activity | Reason for Conducting | Details |
|---|---|---|
| **Area Programme baseline** | 1. Confirm the need for literacy programming.<br>2. Informs technical approach and programme.<br>3. Establishes baseline measurement of: ***proportion of Grade 3 children that can read with comprehension.*** | ***Tool*:** STAR<br>***Target population*:** Grade 3 children<br>***Sampling frame:*** All grade 3 children enrolled in and attending WV supported schools.<br>***Frequency:*** Once during AP development. |
| **Area Programme evaluation** | 1. Measures progress toward programme goals.<br>2. To measure **change over time of: *proportion of Grade 3 children that can read with comprehension.*** | ***Tool*:** STAR<br>***Target Population*:** Grade 3 children<br>***Sampling Frame**:* All grade 3 children enrolled in and attending WV supported schools<br>***Frequency**:* Every 3-5 years or end of strategy cycle. |

**Although data is collected at AP level, analysis and reporting can be done at different levels as data is brought together. Data collected from different APs can be combined to enable cluster/regional level and national TP level analysis.**

**AP level data** **aggregated to** ⟹ **Cluster/Regional level data** **aggregated to** ⟹ **National level TP data**

### Project model and tool

*World Vision International made a strategic decision to focus on Early Grade reading interventions; hence the Global Centre recommended project model for this age-group is* ***Unlock Literacy/Literacy Boost.***

The recommended tool for literacy assessment at primary education level is STAR. It assesses children in Grade 3 across a programme area. This tool is designed to measure students' progress in reading as a result of a specific literacy programme. It should be used with Grade 3 students enrolled in and attending a structured learning environment[13] to measure a level of minimum proficiency in reading comprehension. All reporting of scores should mention the grade assessed for reference *(e.g. 'Only 2% of children in Grade 3 children could read with comprehension').*

### Target Population

Population: **All Grade 3 children enrolled in and attending a structured learning environment included in WV programming.**

### Rationale for sampling at school level

World Vision's current strategic focus for education is on education quality; the quality of teaching and learning. Thus, sampling from within the formal and non-formal schools in an AP provides a rich source of data to better understand the schools WV is working with or should be focusing on, and to use this information as a starting point for further discussion on the root causes of low reading outcomes in the schools.

Our survey area is the entire geographical area of the programme impact area where WV intends to work over the life of the programme cycle. In this area, ensure that all target schools/structured learning environments in the AP that WV intends to engage in the education development process over time, are included in the baseline sample. This will decrease bias in interpreting changes in literacy outcomes over time, since communities and structured learning environments within an AP may choose to engage in education development initiatives at different times, over the life of the programme.[14]

### Area programme sample size calculator

The initial step is to determine the sample size required per AP using the STAR sampling calculator.

*Note: Although WVI has put together a calculator for determining the optimum sample size required for key TP indicators, this does not work well for literacy indicators which require school-based sampling.*

In many national offices(NO), the proportion of children reading with comprehension in LB pilots was under 10% for students at baseline, but after just a year of programming, a 25% increase was a reasonable expectation if programming was done well. Using the LEAP3 calculator, this makes the sample size very small (less than 80 if started at 5% RWC - for example).

In addition with very little national data available on early grade literacy rates, it is almost impossible for NOs to realistically estimate P1 and P2.

For these reasons, the recommended population based calculator (imbedded below) assumes a 50% proportion to maximize the sample size within an AP, to be able to statistically detect change in literacy outcome from baseline to evaluation.

*The calculator is developed from standard statistical formulae used consistently by a number of academic institutions; (see link for reference[15])*

[13] Structured learning environment includes formal and non-formal schools.
[14] Measuring Quality of Learning: World Vision Starting Point-Jacquee Bunnell, Isabelle Carboni, Deborah Loesch–Griffin and Ann Munene.
[15] Creative Research Systems: https://www.surveysystem.com/sscalc.htm

**Sampling Methodology**

Once the sample size required per AP is determined using the sample size calculator described above and imbedded below, consider using one of the following scenarios to proceed with the sample selection process. *Please note that scenario one is the **preferred option** as it assumes that if a school is selected for intervention, then it will be visited throughout the intervention period.*

**Scenario 1: All intervention schools within an AP are included for baseline/evaluation study:**

**Step 1.** Create a list of all schools receiving WV programming (or planning to receive programming in the case of a baseline). For each school, identify the number of boys and girls enrolled in Grade 3.

**Step 2.** Follow the STAR Sampling Tool instructions to fill out the list of schools and students in an AP- see template provided <insert excel tool here>.

STAR Sampling Calculator_ Jan2018.

**Step 3.** Determine how many boys and girls need to be sampled from each school using the proportionate sampling template - see example given in the sampling template. **Please note that the sample size will be adjusted to ensure at least 15 students per school (or all children in the school if they are less than 15) are included in the sample. This will be taken care of by the sampling tool.**

**Step 4.** At the school, randomly select the required number of boys and girls using the student selection process as shown by step 5 example below.

**Step 5. Students' selection using a systematic random sampling technique[16].**

Assuming ***school Y*** has a grade 3 students' population of 220 (120 girls & 100 boys), and we need to randomly select 22 students from the school (10 boys and 12 girls), stratify the grade 3 students into boys and girls' strata and divide each stratum by the number required, to determine the selection interval. For example, for the girls' it will be 120÷12 = 10, the counting interval would be 10. Every 10th girl on the registration list is chosen until a total of 12 girls have been selected. Repeat the same process for boys. The student selection process should generally always be carried out on the day of assessment.

| **Student selection process: To assess 22 students (12 girls and 10 boys) in a school, conduct the following procedure in each structured learning environment:** |
|---|
| 1. Obtain a list of all Grade 3 student in each class.<br>2. Arrange this list by sex of students, girls and boys.<br>3. Count the total number of girls registered in each class and add them up (like Class A=40, Class B=30, Class C=50, Total=120).<br>4. Repeat 3 above for boys in each class (like Class A=40, Class B=30, Class C=30, Total=100).<br>5. Note the total number of boys and girls and keep a record of this for each school visited.<br>6. For the girls, divide the total (120) by 12 (the number of girls to be assessed) – (120÷12) = 10; this is the counting interval for girls in this school.<br>7. For the boys, divide the total (100) by 10 (the number of boys to be assessed) – (100÷10) = 10; this is the counting interval for boys in this school.<br>8. Note this interval (nth number) and keep a record of it for both girls and boys in each school.<br>9. Use these nth numbers as the counting interval to count from the lists and select every '10th' girl, or every 10th boy to be part of the sample. *(If a school is unisex, select the 22 children and keep a record of this for reporting)* |

[16] Sampling intact classes is not recommended as some schools may have streamed student per learning abilities.

10. If a student is absent or refuses to participate, proceed to the next selected child on the list. If that student is absent or refuses, then proceed to the next. It is advised that NOs ensure that a non-response rate of 10% is included in the initial sample size calculator.

**Scenario 2: When not all schools in an AP can be included for baseline/evaluation study - need to select a school sample:**

*This is a less preferred option and should only be considered where option 1 to include all target schools in the survey is almost impossible.*

Sometimes due to logistical and budgetary reasons in a NO, it may become impossible to survey all intervention schools in an AP. In this case, a representative sample of schools and students will be required.

A **proportionate stratified** sample design is recommended to obtain a representative sample of schools and students in an AP.

When background information about the schools is available, use this to determine and categorise the schools in an AP into 2 or 3 groups (called strata), based on similarities or differences in characteristics and/or dimensions. For example, if schools in an AP vary in student population size, stratify schools by size - small, medium and large. Or you can use the physical condition or resources available within the school environment such as: Good, Poor and Very Poor.

If the background information required to categorise schools is not readily available, it is recommended that a situation assessment/analysis on schools be conducted to identify probable characteristics and dimensions to help group the schools.

**Step 1:** Identify all relevant strata given the school characteristics information available, ensuring their actual representation.

**Step 2:** From the list of all schools in an AP, place each school in the most appropriate stratum and number each school within each stratum with a unique identification number.

**Step 3**: Use **proportionate stratified random sampling** technique to determine the sample size for each stratum, and select sufficient numbers of schools from each stratum. This means if the number of grade 3 students in a particular school comprise, for example, 20% of the total number of grade 3 students in all the schools included in the sample, then 20% of the students in the sample should come from this school.

*Note: It is important that school sample from each stratum be selected in a random manner to minimize bias and improve representativeness.*

**Step 4:** Create a list of selected schools per strata and follow **steps 2-5 of scenario 1** to select a proportionate sample of boys and girls from each school selected in step 3 above.

---

In the event that both scenarios illustrated here would **not** work for your contextual situation, please contact GC Evidence & Learning or GC Education sector team for further guidance and support.

---

**Points to remember**:

- ✓ If the programme design and the local context require it, school stratification can be used to ensure inclusion of ALL types of structured learning environments in the sample.
- ✓ In some rural areas or non-formal school settings, there might be very few students enrolled and attending. If a selected school has fewer than, or up to 4 more students than the required number per school, it is important to be aware of any psychological implications on those children not included in the sample. So, one way to avoid such implication is to do the assessment on the remaining children after the sampled ones are completed. The data collected from the additional students should be filtered out during data analysis.

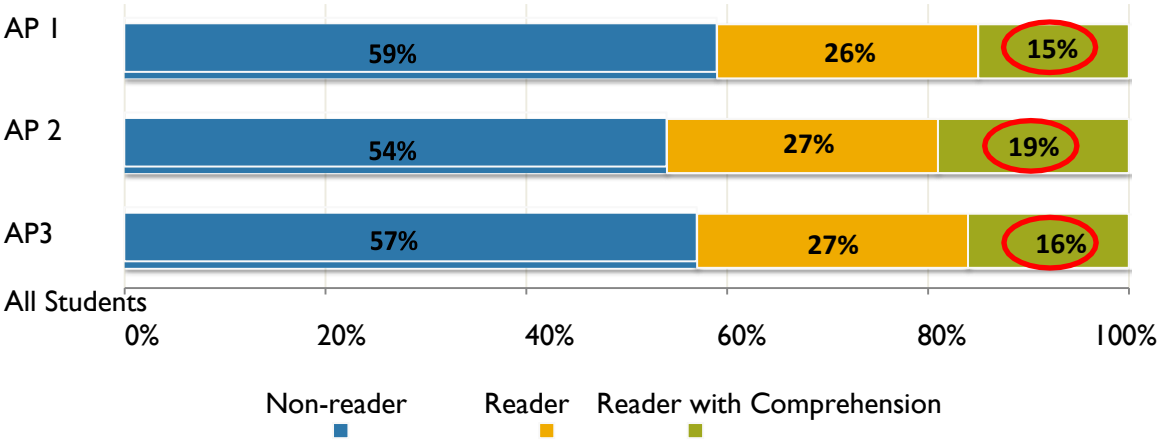
**Reporting for child well-being outcome**[17]

*"Proportion of Grade 3 students who can read with comprehension"*
The STAR tool has different sub-tests. A number of results can be analysed and reported from the tool to provide important information for programming. However, this short guide will only include the reading comprehension sub-test required for CWB reporting (for more detailed analysis and reporting guidance please refer to the full STAR Field Guidance).

Using the example graph given below, to report the "*proportion of Grade 3 children who can read with comprehension",* the NO would report only the figure given in red - **15%** of students from AP 1 are readers with comprehension, **19%** of students from AP 2 are readers with comprehension, and **16%** of students overall are readers with comprehension. These are the students meeting the child wellbeing outcome.

Readers with comprehension = those students who were identified as readers and received a score of 4 or higher (≥ 80%) on the comprehension questions (green bar in the graph below and % circled in red).

**Readers with comprehension tiers by AP**




*Note: These reader 'tiers' can also be calculated according to other groups like sex, ethnic groups etc. (the above chart calculates these scores according to AP)*

## Collecting background and demographic information

There are many factors inside and outside of schools that influence a child's learning. To be able to track and analyse such factors, it is important to always gather student background information during assessment. This can be done through a simple **background questionnaire** administered orally to each student during assessment as part of AP baseline/endline.

[17] STAR guidance (school-based test about reading). *A guide to star tool preparation and contextualization:* World Vision International 2016

*Note: It is also recommended that population level data on **access to education** (enrolment and attendance), for school going children (6-18 years) is collected at household level using the caregiver survey education module. A sample size for the access indicator will be determined at household level using the LEAP3 sample calculator based on the prevalence of enrolment and attendance and the change we want to achieve.*

### Considerations for school-based surveys

When doing school-based surveys (as part of TP baseline or evaluation) the following must be considered carefully:

1. **Baseline timing versus school schedule**: Data collection must be timed according to school cycle to avoid clashing with important school events. Consider the academic year start and end dates, school breaks/holidays and examination timetables to avoid conflict with scheduled school events. Students should not be assessed during holiday breaks, nor should students who are not in attendance on the day of assessment be located at their homes to be assessed. Students should only be assessed at the school during a regular school day.
2. **Evaluate at the same time:** To account for children progressing through the curriculum and breaks in learning due to school holidays, it is essential that the baseline/evaluation assessment(s) take place during the same month in subsequent years.
3. **Timeframe for completing assessment:** Ensure that the data collection in completed in a timely manner. At most, school-based assessment should not stretch over more than 3 weeks. Overstretching the assessment period means that those assessed towards the end will have learnt more skills than those assessed at the beginning, hence the reading results may be skewed.
4. **Consideration for MVC:** It is important to collect socio-demographic information to identify/highlight issues of (in)equity and the needs of the most vulnerable children (MVC). An identifier variable(s) should be included to allow for analysis of different marginalized populations and to ensure those who are disadvantaged can be targeted accordingly. The STAR tool includes background questions to be asked of the students, and when complemented by school environment data collected during a school environment survey, can help NO to target interventions.

*Notes: For more details and further support regarding sampling methods, please contact the MSE DME team or Education sector GC team.*

**Ann Munene**, *Education & Life Skills M&E Senior Advisor: ann_munene@wvi.org*
**Albana Spiro,** *DME Director: albana_spiro@wvi.org*
**Andrew Clucas,** Senior Advisor, Quality and Innovation: Andrew_Clucas@wvi.org

# Annex 6: Application of design weights during surveys

A technical programme (TP) baseline survey may be conducted using the same sample size of households in each Area Programme (AP) for the caregiver survey. Since the population (number of HHs in this case) varies among APs, simply aggregating the samples to generate TP level baseline value will not then give results which represent population values. Therefore, during analysis of the survey data, ensuring that the sample is representative of the population is very important. Application of sample weights helps in this regard.

There are different types of weights. Here only design weights to ensure representation of populations in the various APs included in the survey will be considered. This hypothetical example helps to explain the use of design weights.

## Example - Determining weights

A caregiver survey may be conducted in 5 APs with a sample size of 15 households from each AP. The number of households in each AP and related sample weights are provided as follows.

| Area Programme name | Total number of HHs in the area (T) | #HHs in the sample (S) | Probability of a HH being selected in the sample (P=S/T) | Weight (W = 1/P or W=T/S) |
|---|---|---|---|---|
| **Green** | 310 | 15 | 0.048 | 20.67 |
| **Yellow** | 550 | 15 | 0.027 | 36.67 |
| **Red** | 475 | 15 | 0.032 | 31.67 |
| **White** | 684 | 15 | 0.022 | 45.60 |
| **Blue** | 350 | 15 | 0.043 | 23.33 |

## Generating TP level baseline values

For demonstration purposes, let's take two indicators from the Livelihood TP.

Indicator #1: Proportion of HHs who have year round access to food

Indicator #2: Annual average income of the HH in USD

The summary of baseline results per AP and the TP level averages (un-weighted and weighted) are provided as follows.

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| **Area Programme name** | **Proportion of HHs who have year round access to food (YRA)** | **Annual average income of the HH in USD (AVI)** | **Weighted YRA = YRA x W** | **Weighted AVI = AVI x W** | **Weight (W)** |
| **Green** | 0.56 | 500 | 11.57 | 10,333.33 | 20.67 |
| **Yellow** | 0.90 | 876 | 33.00 | 32,120.00 | 36.67 |
| **Red** | 0.32 | 450 | 10.13 | 14,250.00 | 31.67 |
| **White** | 0.73 | 689 | 33.29 | 31,418.40 | 45.60 |
| **Blue** | 0.90 | 788 | 21.00 | 18,386.67 | 23.33 |
| **Average (TP level)** | **0.682** | **660.60** | **0.690** | **674.39** | **157.93** |

As can be observed from the table above, the TP level un-weighted baseline values are different from the weighted ones.

| Indicators | TP level baseline values | |
|---|---|---|
| | Un-weighted | Weighted |
| **Indicator #1: Proportion of HHs who have year round access to food** | 0.682 (68.2%) | 0.690 (69%) |
| **Indicator #2: Annual average income of the HH in USD** | 660.60 | 674.39 |

**Brief explanation how the averages have been calculated.**
Un-weighted average is the sum of the values divided by the number of values:

For indicator #1: (0.56+0.90+0.32+0.73+0.90)/5 = **0.682**

For indicator #2: (500+876+450+689+788)/5 = **660.60**

Weighted average is the sum of the weighed values divided by the sum of the weights

For indicator #1: (11.57+33.00+10.13+33.29+21.00)/ 157.93 = **0.690**

For indicator #2: (10,333.33+32,120.00+14,250.00+31,418.40+18,386.67)/ 157.93 = **674.39**

## Conclusion

The application of weights improves the representativeness of the sample. The example above shows differences between weighted and un-weighted values. When the variation between the population sizes of APs gets higher, the sampling error gets higher. Therefore the application of sample weights would help minimize the sampling error which could be inflated as a result of such variation in population sizes.

It should also be noted that, depending on the type of indicator, weights can be based on other demographic characteristics such as proportion of people in certain age group (such as youth age 12-18, children 0-5 years…etc), gender or social status (eg. level of education) – though it is subject to availability of such data about the population.