Meta-Evaluation of 33 Evaluation Reports of World Vision Germany

on behalf of World Vision Germany

Author:
Tatjana Mauthofer
Quality Assurance:
Dr. Stefan Silvestrini

Center for Evaluation
Im Stadtwald
Geb. C 5.3
D-66123 Saarbruecken

Phone +49 – (0)6 81 – 3 02 36 79
E-Mail s.silvestrini@ceval.de
t.mauthofer@ceval.de
URL http://www.ceval.de

Saarbrücken, February 23, 2018

Table of contents

Abbreviations 1
List of Figures 1
List of Tables 1
Executive Summary 2
1. Background 5
2. Methodological Approach 6
3. Assessment according to quality of evaluation criteria 8
3.1 Voice and Inclusion 8
3.2 Transparency 9
3.3 Methodology 10
3.4 Triangulation 13
3.5 Contribution 14
3.6 Satisfaction of Information Needs 16
3.7 Conceptualization of Findings 17
3.8 Sustainability 18
4. Conclusion 19
5. Recommendations 20
6. Annex 21
6.1 Criteria, sub-criteria and scale definitions 21
6.2 Graphical illustration of sub-questions 28
6.2.1 Voice and Inclusion 28
6.2.2 Transparency 28
6.2.3 Methodology 29
6.2.4 Triangulation 29
6.2.5 Contribution of WV's interventions 30
6.2.6 Satisfaction of information needs 30
6.2.7 Conceptualization of Findings 31
6.2.8 Sustainability 31
6.3 Summary of Results 32

Abbreviations

ADP	Area Development Program
CEval	Centrum for Evaluation
DAP	Developmental Assets Profile
FLAT	Functional Literacy Assessment Tool
FGD	Focus Group Discussion
MENA	Middle East & North Africa
NGOs	Non-Governmental Organization
ToR	Terms of Reference
WVG	World Vision Germany
WV	World Vision
YHBS	Youth Health Behavior Survey

List of Figures

Figure 1: Summary of Results 2
Figure 2: Overall performance referring to the criterion Voice and Inclusion 8
Figure 3: Overall performance referring to the Transparency criterion 9
Figure 4: Overall performance referring to the Methodology criterion 11
Figure 5: Overall performance referring to the criterion Triangulation 14
Figure 6: Overall performance referring to the criterion Contribution 15
Figure 7: Overall performance referring to the criterion Satisfaction of Information Needs 16
Figure 8: Performance referring to organization of findings 17
Figure 9: Performance referring to Sustainability 18

List of Tables

Table 1: Main Conclusions and Recommendations 4
Table 2: Geographical distribution of evaluation reports 6
Table 3: Rating system applied 7
Table 4: Application of WV's data collection instruments and innovative qualitative methods 12

Executive Summary

After the meta-evaluations in 2014 and 2016, World Vision Germany (WGV) has commissioned the Center for Evaluation (CEval GmbH) for the third time to assess the methodological soundness of their evaluation reports. This year's given sample consists of 33 evaluation reports, which have been produced in the financial years 2016 and 2017. The meta-evaluation team differentiates between Area Development Program (ADP) reports (19) and other project evaluations (14). Reports comprise Light Touch[1] (4), Mid-Term (6) and Final Evaluations (23) and cover 17 countries spread over four world regions. The underlying assessment criteria of the meta-evaluation build on previous studies, aligned to the internationally established Bond Evidence Principles, but have been slightly modified and expanded. They comprise *Voice and Inclusion, Transparency, Methodology, Triangulation, Contribution, Satisfaction of Information Needs, Conceptualization of Findings* and the newly added criterion of *Sustainability*.

In a first step, the existing analysis matrix was revised and slightly modified. It was structured along the eight main criteria, each comprising three to six sub-criteria to cover a range of different aspects of report quality. In line with the 1-4 scale provided by the Bond tool, the research team defined scales for non-Bond questions accordingly. The overall grading system differentiates between the following four categories 1=very poor, 2=poor, 3=fair, and 4=good. Ratings for each sub-question were consolidated and one aggregated rating was calculated for each of the eight criteria. The quality threshold is passed when either receiving a 'fair' or 'good' rating. An additional color code was given for best practices. To further improve on the reliability and transparency of findings specific scale definitions were defined for the first time in the meta-evaluation at hand. Figure 1: Summary of ResultsFigure 1 depicts overall results for each assessment criterion, revealing that satisfying results have been achieved for five criteria (*Transparency, Methodology, Satisfaction of Information Needs, Conceptualization of Findings, Sustainability)*, while there is potential for improvement for the remaining three criteria of *Voice and Inclusion, Triangulation* and *Contribution*.

Figure 1: Summary of Results

[1] A Light Touch evaluation is an interim evaluation format, which builds on less extensive data collection. It emerged while transitioning from World Vision's internal project management approach LEAP 2 to LEAP 3. An alternative evaluation format was necessary due to new methodologies used in LEAP 3.

Insight on each criterion are summarized in the following:

Voice and Inclusion: The majority of reports received an overall 'poor' rating, nevertheless ratings of sub-criteria varied. Rather positive results were achieved for the question on including the voice of stakeholders and beneficiaries, but marginalized groups have not yet been integrated sufficiently in the evaluation process. Similarly, disaggregation of data rarely goes beyond sex and age. A few truly participatory practices could be identified in terms of involving beneficiaries or partners in the evaluation process.

Transparency: Results are promising with 70% of reports receiving an either 'fair' or 'good' overall rating. Main positive aspects comprise transparency on sample size and composition, methods used and limitations declared. Deficiencies were revealed regarding disclosure of assessment criteria and standards. Furthermore, especially Light Touch Evaluations were often subject to lack of objectivity, and hence, weakened credibility.

Methodology: Overall results are quite satisfying with 72% of reports showing 'fair' or 'good' results. Data collection instruments, sample sizes and analysis tools were displayed in a satisfying manner and limitations of the study were outlined. The usage of monitoring data was, however, rarely mentioned explicitly in the evaluation reports and also the explanation of the program's intervention logic requires enforcement. Innovative quantitative and qualitative data collection instruments continue to be applied with declined usage of FLAT, DAP, Tree of Change, among others and emerging usage of the 'Most Significant Change Technique', among others.

Triangulation: Assessments varied tremendously and only 45% received an either 'fair' or 'good' overall rating. In terms of triangulating data, it could be noted positively that different stakeholder perspectives were exposed, however divergent findings within a stakeholder group were not yet contrasted sufficiently and put into perspective appropriately. Furthermore, above all Light Touch Evaluations show more deficits.

Contribution: Results were found to be one of the weakest in comparison to the other criteria assessed, with only one third of reports passing the quality control threshold. Result chains or logic models were not used coherently and it seems as if the benefits they bring along were not clear to the stakeholders involved. While references to baseline data are quite common, poor quality often impedes a sound comparison between the before and after situation. Thus, it remains often unclear how and to which extent the intervention contributed to the observed results.

Satisfaction of Information Needs: Most reports received at least a 'fair' rating, hence passing the quality control. Still, lessons learned were not always included and recommendations could be more specific in some cases.

Conceptualization: Overall results are very positive and especially the sub-criteria on a well-structured executive summary and a coherent report structure stands out in this regard. None of the reports has received a 'very poor' rating.

Sustainability: Being assessed for the first time, the newly added criterion shows promising results, since the majority of reports receive an either 'fair' or 'good' rating and a few best practices could be identified. However, five reports do not analyze *Sustainability* at all.

The meta-evaluation concludes with specific recommendations, responding to the main conclusions of the study, as shown in the subsequent table:

Table 1: Main Conclusions and Recommendations

Main conclusion	Recommendation
Results on the *Voice and Inclusion* criteria have deteriorated between the 2016 and this year's evaluation.	WV should reinforce understanding among project staff that evaluations should be an inclusive and participatory process. The exchange with beneficiaries needs to be promoted during several stages of the evaluation, be it during an inception workshop or during the validation of findings and formulation of recommendations.
Perspectives of different stakeholders are displayed, but conflicting findings within groups are less elaborated on.	ToRs should emphasize the examination of divergent findings and opinions when analyzing qualitative findings, since this does not only enrich evaluation reports but also gives voice to different key stakeholders.
There is low usage of result chain and programmatic logic tools.	WV could promote awareness creation on benefits of analytical tools in specific workshops and anchor the compilation of tools (e.g. Theory of Change) in their ToRs as deliverable of the consultancy.
It is a good intention to use new-age, participatory tools, but actual application must be improved to achieve full benefits.	WV stands out due to their innovative data collection tools, but some of them require improved capacities for both program staff and evaluators. The actual implementation of these tools must be practiced more frequently to ensure accurate implementation in the field. WV should ensure that evaluators / consultants can prove proficient experience in applying them.
The transition from Leap 2 to Leap 3 and the interim solution of implementing Light Touch Evaluation has consequences.	Since the quality of Light Touch Evaluations is substantially weaker across different criteria, it should be contemplated whether they provide sufficient information and quality to serve evaluation purposes. While this is an interim and time-bound issue, it should still be taken into account when planning another transition phase.

1. Background

Meta-evaluations – the evaluation of evaluations - refer to the systematic review of evaluations to assess their quality against established standards and principles. World Vision Germany (WVG) is one of the pioneering German NGOs when it comes to regularly assessing the quality of its international program and project evaluations.

Meta-evaluations can serve to showcase the adherence to the establishment of good practice in evaluation studies. According to Patton (1997), the key motive for conducting meta-evaluations is to ensure an independent and credible review of an evaluation's strengths and weaknesses. In times of more frequent communication with stakeholders and the urge to showcase transparency and accountability, disseminating information on evaluation practices is of utmost importance to maintain credibility. At the same time, meta-evaluations can serve as learning tools to improve the usefulness and utilization of evaluation findings and hence, achieve better performance.

For the third time, following meta-evaluations in 2014 and 2016, WVG commissioned the Center for Evaluation (CEval) to assess the methodological soundness of their evaluation reports. The chosen sample comprised 19 Area Development Program (ADP) reports and 14 project reports, covering four world regions and 17 countries. The underlying assessment criteria build on previous studies but have been slightly modified and expanded. A *Sustainability* criterion has been newly added and sub-questions of the *Appropriateness of Evaluation Methods* criteria are now captured within the *Methodology* criterion. The majority of analysis criteria applied are oriented by the internationally established Bond Evidence Principles.

This results in the following eight assessment criteria[2]:

1. Voice and Inclusion
2. Transparency
3. Methodology
4. Triangulation
5. Identification of WV's Contribution
6. Satisfaction of Information Needs
7. Conceptualization of findings
8. Sustainability

Each criterion is composed of three to six sub-questions. Unlike in the previous study, this year's meta-evaluation refrained from conducting an online-survey, which had aimed at assessing the usefulness and utilization of evaluation findings. To avoid survey-fatigue and gain relevant findings over time, it made sense to not repeat such an activity every year but leave a sound time gap in between before again approaching staff to inquire about evaluation practices and usage.

The report is structured as follows: A brief introduction on the methodological approach is given (Chapter 3), followed by an in-depth analysis and the description of main findings of each criterion (Chapter 4). Chapter 5 elaborates on conclusions and an overview result table, followed by specific recommendations shared in Chapter 6.

[2] An overview of the criteria including the respective sub-questions and scale definitions can be found in Annex 6.1

2. Methodological Approach

Sample Description

The given meta-evaluation is based on 33 reports, which have been compiled in the financial years 2016 and 2017. To the best of CEval's knowledge this is a complete sample of available evaluations conducted in the considered time period. The study team differentiates between ADP[3] reports (19) and other project evaluations (14), covering 17 countries and four world regions. The geographical distribution is displayed in the following table:

Table 2: Geographical distribution of evaluation reports

Africa		**Asia**	
Tanzania	2	Mongolia	3
Ethiopia	2	Bangladesh	4
Kenya	1	Sri Lanka	4
Mozambique	1	Georgia	1
Burundi	1	**Latin-America**	
Uganda	1	Bolivia	5
Zimbabwe	1	Honduras	2
Sierra Leone	1	**Middle East – Eastern Europe**	
Congo	1	Jordan	2
Sudan	1	**TOTAL**	**33**

Considering that the given sample contains Light Touch (4)[4], Mid-Term (6) and Final Evaluations (23), three types of evaluations with very different purposes and resources available, this meta evaluation cannot offer a sound comparison between evaluation reports, but rather assesses their individual quality against the established assessment criteria and shows tendencies of the overall sample.

Data Analysis

The evaluation is based on an analysis matrix, jointly created by CEval and WVG, which is majorly shaped by the internationally renowned Bond Evaluation Principles. The analysis matrix is structured along the main criteria of *Voice and Inclusion, Transparency, Methodology, Triangulation, Contribution, Satisfaction of Information Needs, Conceptualization of Findings* and the newly added criterion of *Sustainability*. The latter builds, to the best extent possible, on WV's 'Five Drivers of Sustainability', which comprise Local Ownership, Partnering, Transformed Relationships, Local and National Advocacy and Household and Family Resilience. Further changes between this year's and previous meta-evaluations center around the *Methodology* criterion, which now captures sub-questions of the formerly used criterion of *Appropriateness of Methodology*. Feedback received revealed that consolidating the two criteria will lead to improved reader-friendliness and better understanding by WV's audience.

[3] An ADP can be understood as a program in a selected district or region (depending on the population density) which comprises usually three to five projects. All ADPs put a strong focus on child well-being and thus have a sponsorship, an education and a health project in common. However, they vary according to projects related to community development, which are often in areas like agricultural development, vocational training or improved water and sanitation. In general, ADPs are running for about 15 years and are evaluated at different points in time.

[4] A Light Touch evaluation is an interim evaluation format, which builds on less extensive data collection. It emerged while transitioning from World Vision's internal project management approach LEAP 2 to LEAP 3. An alternative evaluation format was necessary due to new methodologies used in LEAP 3.

To avoid oversimplification and explore different dimensions within the main categories, there are three to six sub-questions for each of the eight criteria. While the Bond tool[5] already provides a 1-4 scale for each sub-question, the evaluation team defined scales for non-Bond questions accordingly. Specific scale definitions had not been defined in the previous meta-evaluations due to time constraints but were included in the meta-evaluation at hand to improve on the reliability and transparency of findings.

In line with the scale compiled, the overall grading system differentiates between the following four categories with a numerical value between 1 and 4 for each, i.e. 1=very poor, 2=poor, 3=fair, and 4=good. An additional color code is given for best practices. Table 2 describes the rating system:

Table 3: Rating system applied

Numerical Value	Descriptive Rating	Quality Control Decision
1	Very Poor	*Fail*
2	Poor	
3	Fair	*Pass*
4	Good	
4	Best Practice	

Ratings for each sub-question were consolidated and one aggregated rating – the mean – was calculated for each of the eight criteria for further analysis. This evaluation study uses the following terminology for assessing overall quality control: A report is declared to have passed a certain criterion if it either receives a 'fair' or 'good' rating. Contrarily, a report fails the quality control, if it receives an overall 'poor' or 'very poor' rating.

To underpin quantitative results with qualitative examples in the study, the reports were analyzed in the qualitative data analysis software MaxQDA®, highlighting best- and worst-case examples for each criterion.

Limitations

Due to the high number of reports within the sample and limited time provided per document, not every aspect of a report could be assessed in full detail. This study rather aims at providing evidence on de facto application of WV's evaluation guidelines, data collection instruments, application of innovative methods and appropriateness of data interpretation to provide a hint on challenges and to highlight promising evaluation cases.

Furthermore, researcher bias cannot be fully eliminated when qualitative information is quantified. To mitigate this risk, a senior researcher conducted a cross-check of two evaluation reports to assess whether the ratings are similar for both researchers. Adjustments were made accordingly.

Lastly, a comparison of the overall result of this year's document analysis and the previous ones have to be handled with care and can only be interpreted restrictively, since the changes in budget, conditions and guidelines within the timeframe could not be considered within the scope of this study.

[5] Information on the tool can be found here: https://www.bond.org.uk/effectiveness/principles#download

3. Assessment according to quality of evaluation criteria

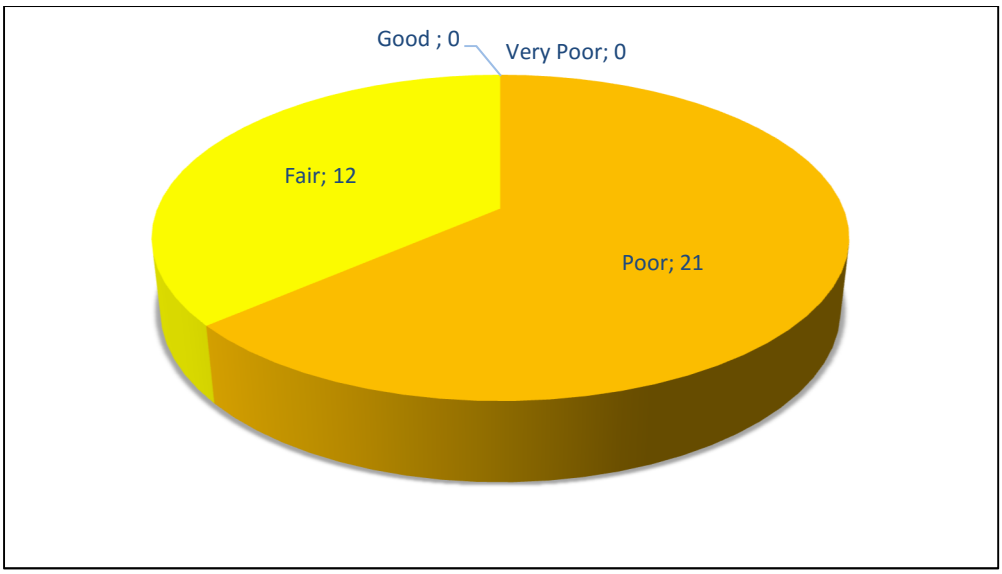
The following section describes findings for each criterion, graphically outlining aggregated results in pie charts. Moreover, specific results for sub-questions are examined and examples for good practices given. In Annex 6.1 the filled analysis matrix can be found.

3.1 Voice and Inclusion

This section focuses on the criteria of *Voice and Inclusion* and highlights to which extent the beneficiaries' views on the effects of the intervention are presented in the evaluation reports. It further shows whether the views of the most excluded and marginalized groups are adequately incorporated and whether findings were appropriately disaggregated according to sex, disability or other causes of social differentiation. Lastly, it is assessed whether the reports clearly identify how and up to which level partners and beneficiaries were included during the interventions of the project.

The aggregated results in Figure 2 show that 21 reports perform poorly and 12 reports are assessed fairly.

Figure 2: Overall performance referring to the criterion Voice and Inclusion



Looking at different sub-criteria, a more heterogeneous picture can be found. The sub-question 'Is the perspective of beneficiaries and stakeholders included in the evidence?' counts with positive results. The majority, i.e. 22 reports, obtains a fair and 7 reports receive a good assessment. This positive assessment is also reflected in the coherent implementation of Focus Group Discussions (FGDs) with a broad range of stakeholders. It is remarkable that FGDs were applied in 30 out of 33 reports.

There is yet room for improvement in terms of including marginalized stakeholders, especially persons with disability or indigenous / tribal communities. The meta-evaluation could not reveal strong evidence for satisfying results in this regard with more than half of the reports showing either very poor or poor results. Similarly, the results on presenting disaggregated findings appropriately are dissatisfying (0 rated 'good', 11 'fair', 11 'poor', 11 'very poor').

Improvements can be acknowledged for the sub-question on beneficiaries playing an active role in designing the evidence gathering and analysis process. The 'SEEC Evaluation, Sri Lanka' report can be emphasized in this regard, since the evaluation team clearly elaborated on their truly participatory approach:

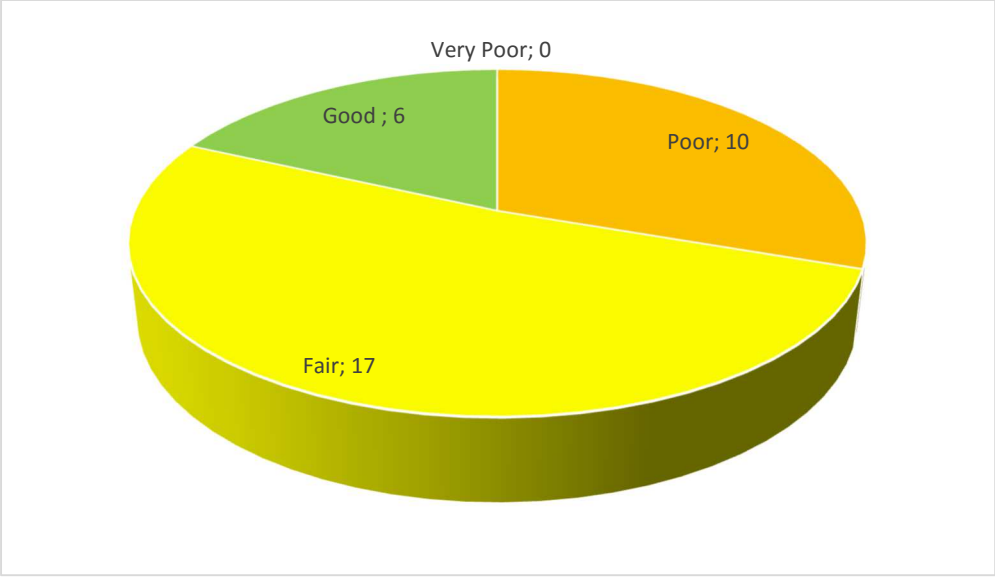
> *"Evaluation purpose was well communicated with stakeholders in advance. Evaluation tools were shared among beneficiaries and stakeholders. Consultative and participatory Interviews, discussions and observations etc. were utilized in this regard. […] Thereafter, findings were presented to selected beneficiaries and for stakeholders. Based on the comments and suggestions validation were completed."*

Lastly, the sub-question on whether the inclusion of beneficiaries and other stakeholders during the interventions was identified in the reports, shows improved results with five out of 33 reports being assessed as good and seven reports assessed as fairly. This resembles an improvement when comparing it to the 2016 results, since in the previous meta-evaluation only two out of 29 reports showed clear evidence on this issue. Results for each sub-question of the *Voice and Inclusion* criteria are displayed in annex 6.2.1.

3.2 Transparency

The transparency of an evaluation is characterized by openness about data sources and methods used, the results achieved, the strengths and limitations of the evidence and the objectivity of evaluators. Aggregated findings on the *Transparency* criteria, depicted in Figure 3, reveal that almost half of the reports perform fairly and six are even rated as good. Ten reports still require improvement, whereas no report has completely failed in this respect.

Figure 3: Overall performance referring to the Transparency criterion



A more detailed assessment of sub-questions reveals satisfying results for transparency on sample size and composition (13 = good, 9= fair, 11= poor) and methods used and limitations declared (14 = good, 11= fair, 8 = poor). In that regard, improvements towards the 2016 study can be recognized, where almost half of the reports had received a poor/ very poor rating. Acceptable results were also achieved when it comes to objectivity of results, since for the majority of the reports (9=good, 18=fair, 6=poor),

an external evaluator led the data collection. Nevertheless, more detailed information on the evaluators' background and institutional affiliation is still missing and should be demanded by WV to support transparency on the objectivity criteria. In the reports which received poor ratings in terms of objectivity, evidence could be found that WV staff accompanied the data collection (e.g. ADP Ephrata, ADP Zuunkharaa) or even conducted the data collection themselves (e.g. ADP Saua Saua Mozambique, ADP Bayankhoshuu). Light Touch Evaluations, most probably due to their nature and resource constraints, are especially subject to lack of objectivity, and hence, weakened credibility.

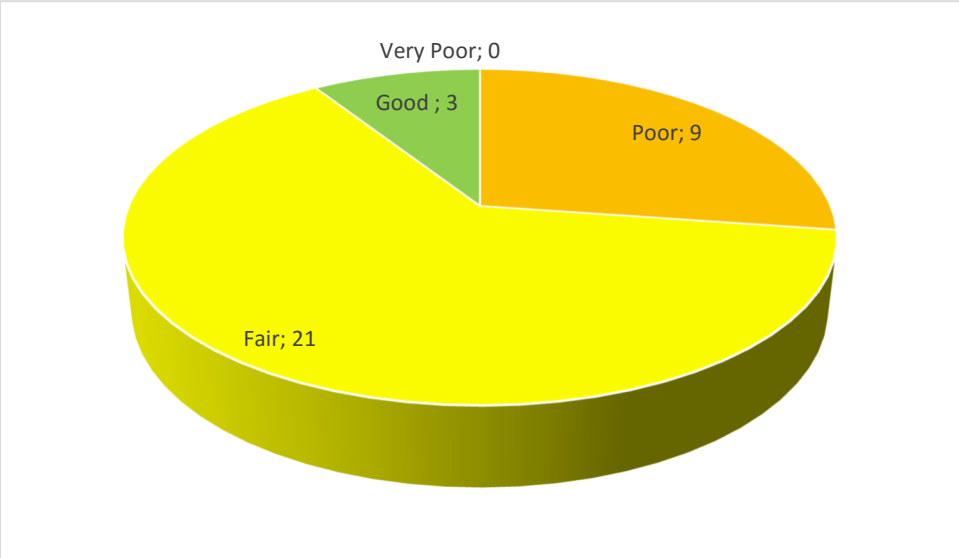
The sub-question on establishing a clear logical link between the conclusions and recommendations presented and the analysis of the collected data shows mixed results with 14 reports are considered 'fair' and 13 reports show 'poor' results. Among the five reports that receive a 'good' rating, the report on the 'Kubum and Mershing Localities Project in Sudan' stands out due to its well-structured overview table which showcases conclusions, backed with findings from the data analysis, and connects them with specific and realistic recommendations. Lastly, deficiencies are revealed in the sub-question on assessment criteria and standards, where only eleven reports perform either 'good' or 'fair' and 22 reports show 'poor' or 'very poor' results. All Light Touch Evaluations show substantial need for improvement in this regard as no evaluation criteria or standards are applied. In contrast, the 'Azraq Camp Project report' in Jordan can be considered as best practice, as DAC criteria are not only listed but also backed up with context specific evaluation questions.

3.3 Methodology

The methodology criterion embraces various dimensions concerned with methodological soundness of evaluation studies, such as relevance of data collection methods and collected data, validity of conclusions in respect to sample, a well-articulated result chain or logic model as well as specific information on data collection, analysis, sampling and limitations.

The overall results, aggregating the six sub-questions, are satisfying with 21 reports showing fair results and three reports showing good results (see Figure 4). Still nine reports could improve on their methodological robustness. Again, this resembles an improvement towards the 2016 studies, where 19 out of 29 reports did not comply with this quality criteria showing poor or very poor results.

Figure 4: Overall performance referring to the Methodology criterion



Considering the scope of these quality criteria, it is of interest to examine sub-questions more in detail. Good results have been achieved for the sub-questions on relevant data collection methods/ reliable data (14=good, 14=fair, 5=poor, 0=very poor), relevant and appropriate data (17=good, 15=fair, 1=poor, 0=very poor) and specific data collection, sampling and analysis methods, data sources and underlying limitations (15=good, 10=fair, 8 =poor, 0 =very poor).

The answers to the question on valid and appropriate conclusions with regard to sampling and sample size are quite varied. Three reports show good and ten show fair results. However, 15 are assessed poorly and five very poorly. It could be identified that the issue is that, most of the time, robust sampling strategies are conducted, however, the sample size and adequateness is no longer referred to when elaborating on conclusions and recommendations.

A weakness can still be seen in the articulation of a result chain or LogFrame. 14 reports do not display either of the two, being ranked 'very poor' and an additional six receive a 'poor' rating. However, at least there has been a positive trend when comparing it to the 2016 study. Whereas 85% of reports were rated as either 'Poor' or 'Very Poor' in 2016, in the given study only 60% received a dissatisfying rating. Moreover, some best practices can be recognized, such as the 'The Mamanieva Project' in Sierra Leone, where a graphical representation of the theory of change is backed with an explanative description of half a page. Observing the low usage, it seems as if there was a lack of awareness on the benefits of articulating on how change occurs. Result chains or logical frameworks are tools that do not only facilitate the evaluation process itself, but also promote better understanding of program activities among staff and external stakeholders.

The re-defined sub-question on 'Monitoring data' received an overall low rating, driving down the aggregated rating of the criteria. These results resemble the ones of the 2016 study. A potential explanation could be that this is either not asked for in the Terms of References (ToRs) or not explicitly mentioned in the reports despite being used. WVG should discuss this issue with CEval to re-assess the relevance of this sub-question, as it majorly impacts the overall rating of the *Methodology* criteria.

Following the approach of the previous studies, the application of quantitative and qualitative data collection instruments shall be examined. The assessment confirms that WV instruments are continued to be used (See Table 3). However, the application has decreased for FALT, DAP, YHBS, Measuring child growth, Seed assessment, Tree of Change between the 2016 study and the one at hand. The application of the caregiver survey, Photo voice, Ladder of Life remain quite stable. One crucial factor influencing the results are the occurrence of Light Touch Evaluations, which are mainly based on traditional qualitative instruments, such as FGDs, KIIs and document reviews. None of the four Light Touch Evaluations makes use of the more innovative instruments promoted by WV and listed in table 3. It can also be observed that new instruments are emerging. In the Latin-American context the child development tools Escala Nelson Ortiz and FELSA are used quite frequently (4 out of 7 Latin-American reports). A Coping Strategy Index was used in the 'ADP Antsokia Gemza, Ethopia' report. In terms of qualitative tools, the 'Most Significant Change Technique' was attempted to be used in three Sri Lankan reports and one report from Uganda. A drawing exercise for children was applied in the ADP Ephrata study and a spider diagram on changes in their community was implemented with teenagers in the ADP Faridpur evaluation.

Comparison between the years must not be overinterpreted, as this study did not control for the subject matter and origin of the reports, as well as for resources provided and objectives set in the ToRs. Still, it is interesting to see that new tools emerge, even though their application might not yet be fully adequate. The 'Most Significant Change Technique' is a quite complex and time-consuming tool, which is based on strong participation of both beneficiaries and management. Photovoice is an interesting approach to gather evidence through photographs taken by the beneficiaries, but the report claiming to have used it, rather provided photographical evidence taken by the researchers. Conclusively, further capacity building might be necessary for WV staff to steer such innovative and truly participatory but complex evaluation processes in the foreseen manner.

Table 4: Application of WV's data collection instruments and innovative qualitative methods

Application of WV's data collection instruments	Financial years 2012 & 2013 N=34		Financial years 2014 & 2015 N=29		Financial years 2015, 2016, 2017 N=33	
	N	In%	N	In%	N	In %
Functional Assessment of Literacy Tool (FLAT)	5	15%	21	72%	7	21%
Development Asset Profile (DAP)	3	8%	7	24%	5	15%
Caregiver Survey	6	18%	14	48%	14	43%
Youth Health Behavior Survey (YHBS)	5	15%	11	38%	6	18%
Measuring child growth	8	24%	9	31%	5	15%
Escala Nelson Ortiz (Latin America)	-	-	-	-	4	12%
FELSA (Latin America)	-	-	-	-	4	12%
Coping Strategy Index (CSI)	-	-	-	-	1	3%

Application of innovative qualitative methods

Comparison discussion group	4	12%	0	0%	1	3%
Photovoice / vision	2	6%	1	3%	1	3%
Seed assessment	5	15%	2	7%	0	0%
Ladder of life	3	8%	4	14%	5	15%
Tree of change	7	21%	7	24%	3	9%
Most Significant Change	-	-	-	-	4	12%
Drawing Exercise	-	-	-	-	1	3%

FGDs (mentioned in 30 reports), observations (mentioned in 12 reports) or key informant interviews (mentioned in 28 reports) are used on a regular basis. Document reviews have been explicitly mentioned in 20 reports.

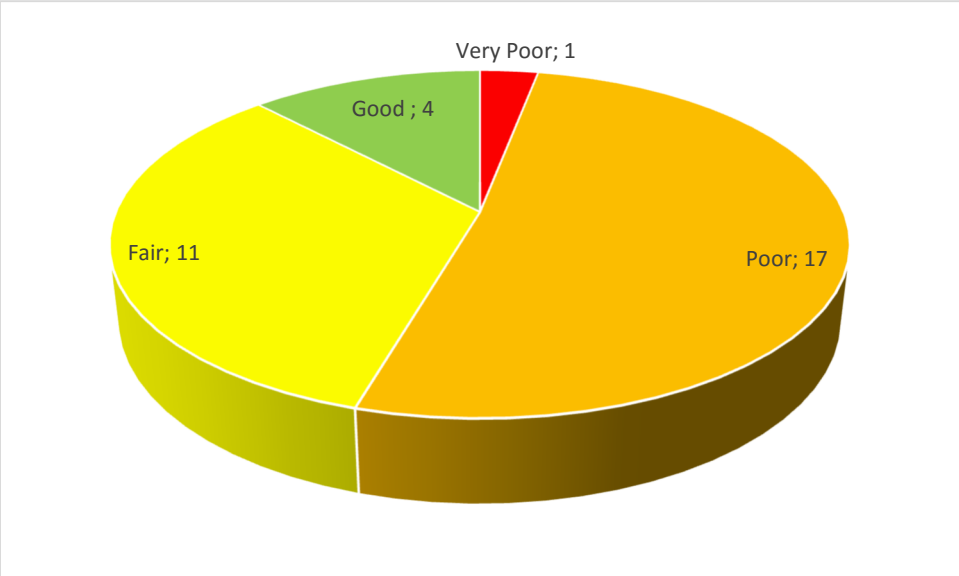
A very positive phenomenon regarding data collection could be retrieved from a few reports of the given sample. The usage of mobile or tablet devices for quantitative data collection has been mentioned in at least two reports. Technology can yield many benefits for data collection processes, as there are now programs that enable clean, more accurate and fast offline data collection which can directly be transferred to computers for data analysis purposes. The 'Kubum and Mershing Localities Project, Sudan' reports explains the process as follows:

"Quantitative data was collected on Forcier's smartphones utilizing the ODK-based ONA software, an innovative mobile data collection technology on the Android Operating System. ONA software can support multi-language surveys allowing it to be customized to the needs of diverse audiences and environments."

3.4 Triangulation

This section shows to which extent the evaluations use a mix of methods, data sources, and perspectives. According to Figure 5 the overall picture for WV's triangulation has scope for improvement. While even less than half of the reports comply with WV's standards (4 rated 'good', 11 'fair'), 18 do not meet the expectations and require improvement (17 rated 'poor', 1 'very poor'). Results have slightly deteriorated compared to findings of the 2016 study.

Figure 5: Overall performance referring to the criterion Triangulation



A detailed look at the different dimensions of the criteria, however, provides a more positive picture. The sub-question on 'making use of different sources or applying different methods' reveals satisfying results for 78% of the reports with 11 being rated 'good' and 15 'fair'. Similarly, the sub-question on comparing the perspectives of different stakeholders is rated positively for 57% of the reports with 6 being rated 'good' and 13 'fair'. While perspectives of different stakeholders are displayed for more than half of the reports, there is a substantial lack of presenting and explaining diverging and conflicting perspectives of or within different stakeholder groups. Accordingly, almost three out of four reports fail the assessment in this regard (17='very poor', 7='poor'). A critical assessment of different perspectives is needed to holistically evaluate how and why changes come about and what are impeding factors to program success. Yet a differentiating view on stakeholder perspectives is missing in the reports subject to this analysis.

Once again, given the broad heterogeneity among the reports, also good examples can be identified, such as the report on the 'Azraq Camp Project' in Jordan, where different views are described and contrasted in the conclusion:

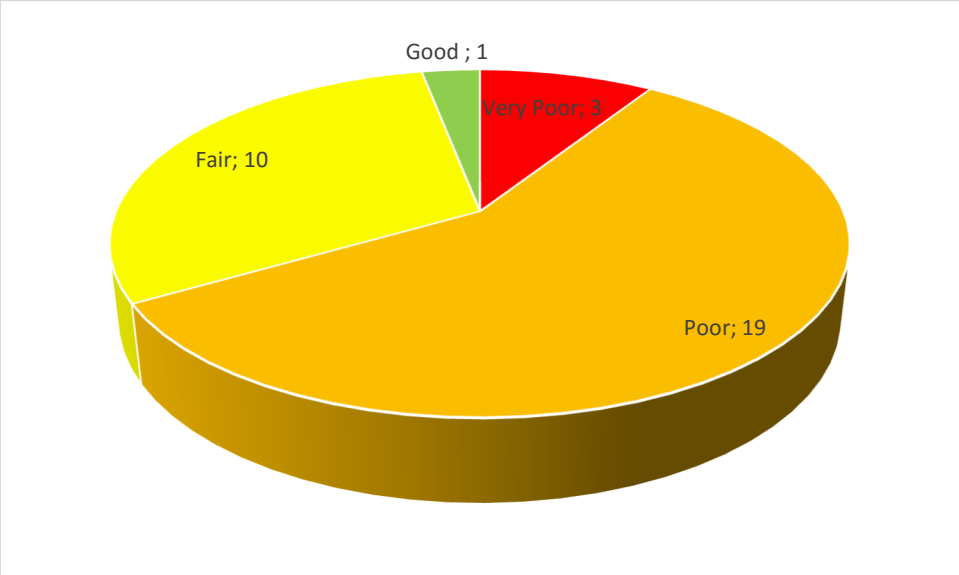
> *"In FGDs with secondary school females, most FGD participants reported that the appearance and taste were good, but the quality and quantity was low, with some referring to "stones" in the date bar, and dryness of the date bar. However, in FGDs with primary school females, most stated they did not like the packaging, but the quality and quantity is good, with one FGD member reporting finding threads in date bar and one reporting finding "stones."*

3.5 Contribution

This section illustrates to which extent the evaluation reports identify WV's contribution to observed changes and what evidence can be found for linking outputs with outcomes. It further examines the acknowledgment of contribution by other actors as well as potential unintended impacts that have occurred.

Figure 6 shows that 19 reports still perform poorly in this regard with three reports even showing very poor results. The remaining 11 have passed the quality control and show either 'fair' (10) or 'good' (1) results.

Figure 6: Overall performance referring to the criterion Contribution



Considering the rather negative result of the aggregated rating, a more in-depth analysis of different dimensions is required to identify main weaknesses. The first sub-question refers to a point of comparison used. Quantitatively this could be a sound baseline study conducted at the beginning of a project or a comparison group, representing a counterfactual to the group that benefitted from the intervention. The latter could also be based on qualitative data collection tools.

Indeed, 25 out of 33 reports refer to baseline data and compare results before and after the intervention. Four reports refer to a comparison group. Accordingly, 8 reports are assessed "good", 13 receive a "fair" rating and the remaining 12 are either 'poor' or 'very poor'. The usage of reference points has improved between the study at hand and the one from 2016. However, there are still weaknesses in terms of scientific rigor. A limitation mentioned several times was that the baseline data was not complete, missing out on several crucial indicators or was not representative. Details on conditions for choosing a comparison group are not revealed to a satisfying extent, which affects the reliability of the results. Statistical testing using SPSS has been implemented by the evaluation team of 'The Mamanieva Project' in Sierra Leone as well as the Sri Lankan ECCD Project, however these remain exceptions.

The sub-question on explaining how the interventions contribute to change shows more balanced results across the four rating categories. 14 reports have passed this quality criterion (6=good, 8=fair) and 19 reports have rather failed in this regard (8=poor, 11=very poor). The 'ADP Natun Jiboner Asha' Bangladesh report shows a profound contribution analysis to explore causal links and can hence be seen as best practice. Some reports, such as the ADP Antsokia Gemza evaluation, analyse causal links implicitly in their 'impact' chapter without sharing a clear overview on the program's logic.

Other sub-questions are rated rather negatively though. Deficiencies are revealed when it comes to 'alternative factors' (2=good, 6=fair, 12= poor, 13= very poor) and 'unintended and unexpected impacts' (0=good, 2=fair, 7= poor, 24 = very poor). Despite being usually mentioned in the ToRs, only few

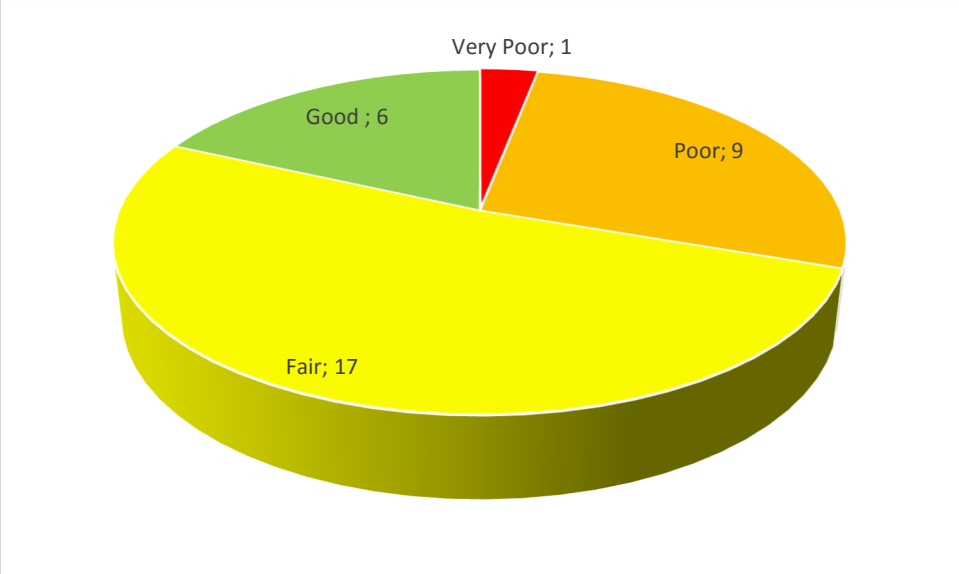
evaluation teams consider these two questions to be important. There has been no improvement between the 2016 meta-evaluation and the current one, emphasizing the need to change this attitude. Enforced awareness raising and capacity building might be necessary.

3.6 Satisfaction of Information Needs

The criterion *Satisfaction of Information Needs* inquires to which extent the evaluation questions in the report are clearly answered and whether a particular evaluation provides WVG with appropriate recommendations and lessons learned.

The overall assessment as displayed in Figure 7 is similar to the study in 2016. Although 10 reports cannot satisfy WVG's information needs (9 rated poor, 1 rated very poor), 17 reports provide a fair and six a good amount of information needed.

Figure 7: Overall performance referring to the criterion Satisfaction of Information Needs



A very good result is achieved for the sub-question on recommendations, with 78% of the reports receive either a good (10) or fair grading (16). The 'Kubum and Mershing Localities, Sudan' report can be regarded as best practice example, as very specific recommendations are given according to the conclusions made. Furthermore, best practices of the programme are mentioned that should be continued to ensure programmatic success.

In can be noted, that the term 'Lessons Learned' is interpreted differently by evaluation teams. Either learnings from the evaluation process are declared or learnings regarding the intervention itself are displayed. Usually, the development community refers to the latter when asking for Lessons Learned. However, reflections on the evaluation process were considered relevant in this study as well and, therefore, received a satisfying rating. In sum, 13 reports were rated as 'good', seven were rated as 'fair' and seven, respectively five, were rated as 'poor', respectively 'very poor'. Among those who received a 'good' rating, most reports elaborated on valuable lessons learned regarding the intervention itself to emphasize certain factors of success.

The question on 'answering the evaluation question has been answered in an accepting manner with room for improvement. 54% of the reports have passed this quality criterion. The 'ADP Antsokia

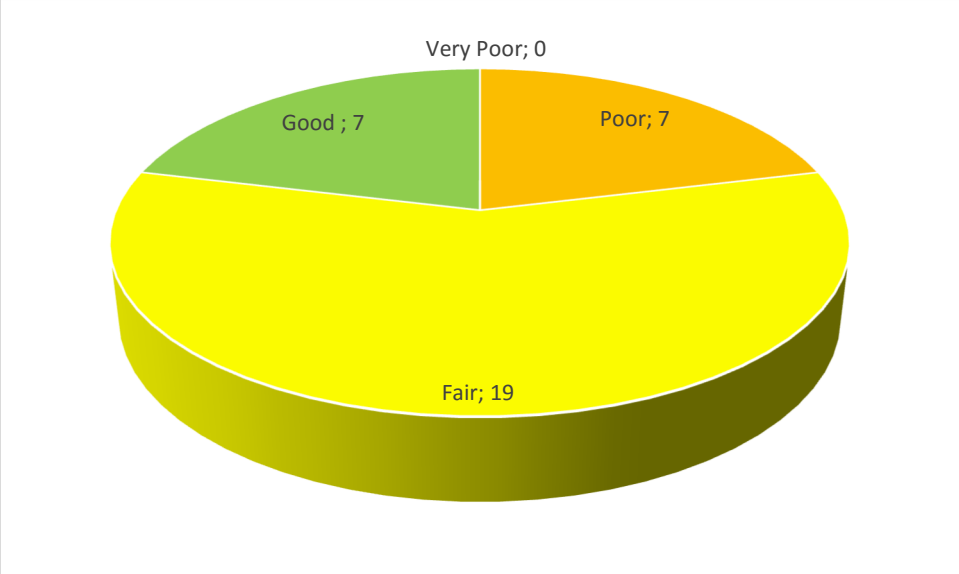
Gemza, Ethopia' report can be highlighted, since evaluation questions are structured along DAC criteria and results are summarized for each criterion at the end of the study.

3.7 Conceptualization of Findings

An easy to follow structure, is not only a precondition to allow outsiders to the project an understanding of the evaluation results, it is also helpful for insiders to capture important findings, most serious limitations and the greater context of the evaluation at first glance.

The aggregated criteria results, shown in Figure 8, look very promising: 78% of the reports show either good (7 reports) or fair (19 reports) assessments. Seven reports are rated poorly, but no report fails completely. There is a slight improvement in quality for these criteria in respect to the 2016 study, in which 62% had passed the quality control.

Figure 8: Performance referring to organization of findings



Remarkably, 70% of the reports have a well-structured executive summary, receiving either a 'fair' (11 reports) or 'good' (12 reports) rating. According to this result, providing an appropriate executive summary has been established as good practice among WV evaluation studies. Almost the same yields true for following a coherent report structure, which has been applied successfully by almost two thirds of the reports. Still 13 require improvement when conceptualizing their findings along well-established evaluation criteria or specific evaluation questions. Acceptable results, with room for improvement, were achieved for the sub-question on 'findings structured along log-frame indicators' (6=good, 9=fair, 14=poor, 4=very poor) and the one on "appropriateness for stakeholders" (5=good, 18=fair, 10=poor, 0=very poor).

Following the positive overall rating, a few best practices can be identified: The 'ECCD Project Sri Lanka' report shows a very convincing structure and a summarizing table of conclusions and recommendations. The 'Kubum and Mershing Localities, Sudan' project report can be distinguished due to its reader-friendliness, supporting photographs and sound balance between figures and descriptive paragraphs.

3.8 Sustainability

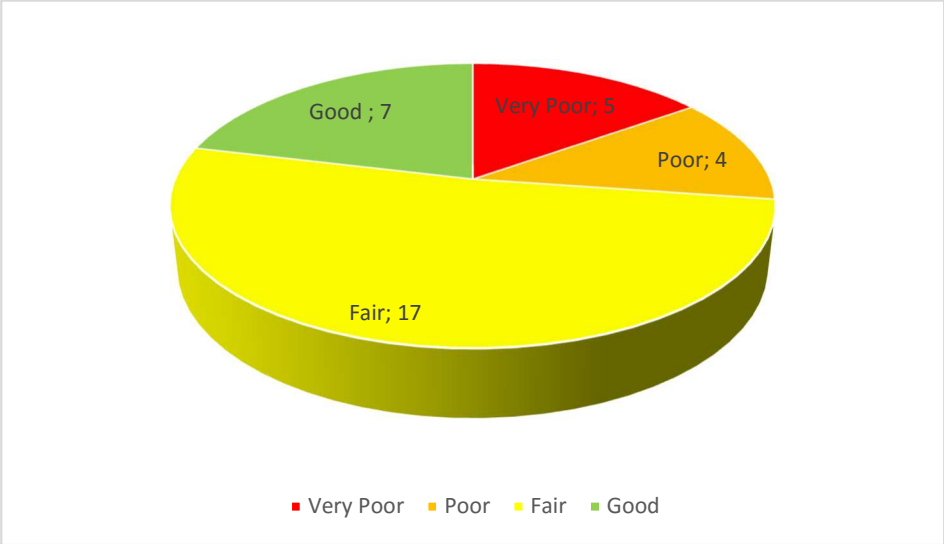
The *Sustainability* criterion has been newly added to the analysis matrix. By doing so, WVG follows the international discourse majorly shaped by the DAC criteria, according to which sustainability is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn (OECD, 2017b). In this evaluation study, the understanding of sustainability is based on WV's paper on 'Five Drivers of Sustainability', which defines it as "the ability to maintain and improve upon the outcomes and goals achieved with external support after that support has ended". It has been an important step forward to add the sustainability dimension to the criteria checklist, and hence to call for a coherent integration of project/program sustainability within WV's evaluation studies. The sub-questions defined for this criterion embrace the reflection of perspectives of beneficiaries to continue with the intervention, ownership and capabilities of partner organization and the overall long-term perspective of the intervention as well as key influencers.

The results are quite varied, as depicted in Figure 9. While seven reports received a good assessment and 17 reports are, at least, rated 'fair', nine reports show deficiencies when reporting on sustainability issues. Five reports do not analyze the criteria at all and, accordingly, receive a 'very poor' rating.

Exploring the sub-questions shows similar tendencies than the aggregated finding on the criteria. 20, respectively 21, reports pass the quality control when it comes to elaborating on stakeholders' perspectives, respectively partner capabilities. 22 reports show either 'good' or 'fair' results regarding the long-term perspective of the intervention.

Considering that the sustainability criterion has been included for the first time, results are indeed very promising. Within the report sample, several best practices could be identified: The "Enhanced Civil Participation Project, Georgia" report describes improved capacity-building, relationships, the produce of a Need Assessment Tool, willingness of local authorities, and ownership towards project results as key drivers of sustainability and shares specific recommendations to promote more sustainable project results. The SEEC, Sri Lanka Project, explores each dimension of the five WV drivers in detail, backing it with relevant quantitative and qualitative data.

Figure 9: Performance referring to Sustainability



4. Conclusion

The overall assessment of the evaluation reports against WVG's evaluation quality criteria, shown in Table 2, summarizes the above results.

The *Voice and Inclusion* criterion does not show a single 'very poor' rating (red color code), but the number of 'poor' ratings (orange color code) needs to be taken seriously to achieve more inclusive evaluation processes in the future. Indeed, the evidence is often grounded in the voice of beneficiaries, but marginalized stakeholders, such as persons with disabilities or indigenous / tribal groups, do not often have a voice in the evaluations. It is positive to see that data is more often disintegrated according to sex and age, but further social differences are not considered when displaying data.

The *Transparency, Methodology* and *Conceptualization* criteria show sound results with no report completely failing and the majority being assessed fairly or even receiving a good rating. In terms of transparent results, 70% passed the quality control. This is a positive take-away, as methodological soundness received a satisfying rating with 72% of reports passing this criterion. The majority of the reports, displayed their data collection instruments, sample sizes and analysis tools, and also outlined limitations of the study in a satisfying manner. The usage of monitoring data is barely mentioned explicitly in the evaluation reports and also the explanation of the program's logic requires enforcement. In addition, evaluation teams begin to use evaluation criteria, such as the DAC criteria, and structure their reports accordingly. 78% of the reports receive a satisfying rating for the *Conceptualization* criteria.

It can be recognized that the criterion of *Contribution* requires improvement, since only a third of reports passed this criterion. The culture of using of tools like result chains or logic models has not yet been established and it seems as if the benefits they bring along are not clear to stakeholders involved. While baseline data is quite common now, the quality of the data is still poor sometimes, impeding a sound comparison between the before and after situation. This finding is in line with the 2016 results, where *Contribution* was found to be a substantial weakness as well.

The *Triangulation* and *Satisfaction of Information Needs* criteria received very varied assessments. In terms of triangulating data, it could be noted positively that different stakeholder perspectives are exposed, however divergent findings within a stakeholder group, are barely examined. This is important, though, since weaker voices might not be considered in the evaluation. Moreover, unlike the rest of the evaluations in the sample, Light Touch Evaluations show deficits in the triangulation of data, since they are merely based on qualitative interviews and/or FGDs. Consequently, only 45% of the reports pass. Looking at the *Satisfaction of Information Needs*, the results do look acceptable since 70% of the reports pass this criterion. Still, lessons learned are not always included and recommendations could be more specific in some cases.

The newly added *Sustainability* shows promising results, since 72% passed the quality control. Nevertheless, considering that five reports showed 'very poor' results, sustainability as an integral criterion of an evaluation study should be enforced in the ToRs and among project staff to ensure coherent usage across all country offices.

Annex 6.3 depicts overall results in a bar chart. Further analyses were conducted to explore differences across regions and change over financial years, however, systematic differences cannot be recognized.

This year's results of the meta-evaluation feature both similarities and differences towards the 2016 study. With regard to the identification of WV's contribution, the issue on quality of baseline data still counts for this year's sample, but overall the use of baseline data has increased again. A lack of comparison groups[6] is still prevalent, but, at least, two reports do show sound use of comparison groups with adequate statistical analysis. Lastly, missing result chains or LogFrames hamper the transparency and understanding causal linkages. Unfortunately, the criterion of *Voice and Inclusion* seems to have deteriorated compared to the 2016 study. Reasons for the same need to be explored in more detail looking at the sub-questions that received unsatisfying overall ratings. In contrast, methodological soundness appears to have improved when comparing it to the last study. This is a pleasant result, hinting that methodological capacities seem to be improving within the focus regions.

5. Recommendations

To further shape its evaluation studies and continue to adhere to good practices in evaluation, WV should consider the following recommendations, structured along selective conclusions.

Main conclusion	Recommendation
Results on the *Voice and Inclusion* criteria have deteriorated between the 2016 and this year's evaluation.	WV should reinforce understanding among project staff that evaluations should be an inclusive and participatory process. The exchange with beneficiaries needs to be promoted during several stages of the evaluation, be it during an inception workshop or during the validation of findings and formulation of recommendations.
Perspectives of different stakeholders are displayed, but conflicting findings within groups are less elaborated on.	ToRs should emphasize the examination of divergent findings and opinions when analyzing qualitative findings, since this does not only enrich evaluation reports but also gives voice to different key stakeholders.
There is low usage of result chain and programmatic logic tools.	WV could promote awareness creation on benefits of analytical tools in specific workshops and anchor the compilation of tools (e.g. Theory of Change) in their ToRs as deliverable of the consultancy.
It is a good intention to use new-age, participatory tools, but actual application must be improved to achieve full benefits.	WV stands out due to their innovative data collection tools, but some of them require improved capacities for both program staff and evaluators. The actual implementation of these tools must be practiced more frequently to ensure accurate implementation in the field. WV should ensure that evaluators / consultants can prove proficient experience in applying them.
The transition from Leap 2 to Leap 3 and the interim solution of implementing Light Touch Evaluation has consequences.	Since the quality of Light Touch Evaluations is substantially weaker across different criteria, it should be contemplated whether they provide sufficient information and quality to serve evaluation purposes. While this is an interim and time-bound issue, it should still be taken into account when planning another transition phase.

[6] The BOND-tool only requests "a point of comparison to show that change has happened (e.g. baseline, a counterfactual, comparison with a similar group)" and hence, not specifically control group and baseline data at the same time.

6. Annex

6.1 Criteria, sub-criteria and scale definitions

Note: The subcriteria colored in grey correspond with those from the BOND tool.		**Scale Definitions**			
Criteria	Sub- Criteria WV Metaevaluation 2017	1	2	3	4
1) Voice and Inclusion We present beneficiaries' views on the effects of the intervention, their participation and identify who has been affected and how	1a. Is the perspective of beneficiaries and stakeholders included in the evidence?	No beneficiary perspectives presented	Beneficariy perspectives presented, but not integrated into analysis	Beneficiary perspectives presented and integrated into analysis	Beneficiary perspectives presented and integrated into analysis, and beneficiaries have validated the findings; the evidence is strongly grounded in the voices of the poor
	1b. Are the perspectives of the most excluded and marginalised groups included in the evidence?	Perspectives from most excluded groups not presented clearly	Perspectives from most excluded groups presented clearly, but not integrated into analysis	Perspectives from most excluded groups presented cleary and integrated into analysis	Perspectives from most excluded presented clearly and integrated into analysis, and excluded groups have validated the findings; the evidence is strongly grounded in the voices of the most excluded
	1c. Are findings disaggregated according to sex, disability and other relevant social differences?	No disaggregation of findings by social differences	Findings are disaggregated, but a number of social differences relevant to the intervention are missing.	Findings are disaggregated according to all social differences relevant to the intervention	Findings are disaggregated according to all social differences relevant to the intervention, and why these have been chosen have been clearly explained
	1d. Did beneficiaries and/or partner organisations play an active role in	Beneficiaries and/or partner organizations	Beneficaries and/or partner organizations actively participated in	Beneficiaries and/or partner organizations had involvement in	Beneficiaries and/or partner organizations had involvement in all of the following

	designing the evidence gathering and analysis process?	have not been involved in the designing and evidence gathering nor in the analysis process	the process and had involvement in <u>one</u> of the following: (1) Designing the process (2) analysing the data (3) formulating the conclusions	<u>two</u> of the following: (1) Designing the process (2) analysing the data (3) formulating the conclusions	(1) designing the process (2) analysing the data (3)formulating the conclusions
	1e.Is clearly identified how and up to which level partners and beneficiaries were included during the interventions of the project?	Only the sectors where interventions took place are mentioned	Incomplete information is given regarding the inclusion of partners and beneficiaries during the intervention of the project	Some interventions are explained and partially connected to beneficiaries	interventions are well explained and connected to the different beneficiary groups
2) Transparency We are open about the data sources and methods used, the results achieved, and the strregths and limitations of the evidence	2a. Is the size and composition of the group from which data is being collected explained and justified?	Size and composition of sample are not described	Size and composition of sample are described	Size and composition of sample are described and justified	Size and composition of sample are described and justified, and all limitations are disclosed
	2b. Are the methods used to collect and analyse data and any limitations of the quality of the data and collection methodology explained and justified?	Methods for data collection and analysis are inadequately described	Methods for data collection and analysis are described	Methods for data collection and analysis are described and justified	Methods for data collection and analysis are described and justified, and all limitations are disclosed
	2c. Is there a clear logical link between the conlusions and recommendations presented and the analysis of the collected data?	Conclusions do not follow from the data collected	Conclusions follow from the data collected	Conclusions follow from the data collected and the steps linking them are clearly explained	Conclusion follow from the data collected and the steps linking them are clearly explained; analysis is transparent about limitations of conclusions

	2d. Are assessment criteria and standards to answer the evaluation questions clearly displayed?	No evaluation criteria or standards are shown	Evaluation criteria or standards are cryptically mentioned: 1-2 sentences	Evaluation criteria or standards are listed but not further explained	All evaluation criteria or standards are shown and explained
	2e. Do the evaluation reports reveal evidence for objectivity or lack of objectivity of the evaluators?	There is strong evidence for lack of objectivity.	There is some evidence for lack of objectivity.	There is some evidence for objectivity.	There is strong evidence for objectiivity.
3) Methodology	3a. Are the data collection methods relevant to the purpose of the assessment and do they generate reliable data?	The methods of data collection are not relevant to the purpose of the assessment and / or the data is unreliable	The methods of data collection are relevant to the purpose of the assessment, but there is uncertainty about the reliability of some of the data	Methods of data collection are relevant to the purpose of the assessment and generate reliable data	Methods of data collection are relevant to the purpose of the assessment and generate highly reliable data; there has been appropriate quality control of the data (eg spot checks, training data collectors)
	3b. Is the collected data relevant and appropritate?	Collected data is not related to project interventions, transformational development cannot be identified	Some data is related to project interventions	Most data is related to project interventions	collected data fully covers the information of interest (reflect relation to the interventions, TDI)
	3c. Are the conclusions valid and appropriate with regard to sampling and sample size?	sample size is not at all considerd in data analysis	Data analysis refers to sample size at one point in the report without further justifying ist validity	Data analysis refers to sample size several times in the report	data analysis cleary refers to sample size; in case a sample or subsample is not representative consequences for data analysis are explained
	3e. Is the results chain or logic well articulated and is the underlying model clearly explained?	no results chain has been detected	the logic of the intervention is briefly summarized	The result chain is explained, but not integrated in analysis.	results chain is (graphically) outlined and explained, results chain is integrated in the analysis

	3f. Does the report specify data collection, sampling and analysis methods, data sources and underlying limitations?	no details on data collection process, sampling and analysis methods nor implicit limitations are outlined	Insufficient details on data collection process, sampling and analysis methods are revealed and information on limitations remain unsatisfactory	Information on data collection process, sampling and analysis methods is given to some extent and a few limitations are mentioned	The sampling strategy is outlined very well, data collection process and analysis as well as limitations are explained appropriately.
	3h. Is monitoring data used and analysed in the evaluation report?	No monitoring data is reviewed or used	the existence of monitoring data was mentioned , but not further used	Monitoring data is reviewed and explained, but not further analyzed	Monitoring data is used and analyzed in the report, brought in line with the intervention's logic / result chain
4) Triangulation We make conclusions about the intervention's effects by using a mix of methods, data sources, and perspectives	4a. To what extent was data triangulated by making use of different sources or applying different methods?	Only one data collection method and source of information is used	Two data collection methods and sources of information are used	Three data collection methods and sources of information are used	Three or more complimentary and distinct data collection methods and sources of information are used and the triangulation of data is explained
	4b. Are the perspectives of different stakeholders compared and analysed in establishing if and how change has occurred?	Different steakholder perspectives have not been presented	Different stakeholder perspectives have been presented but not analysed	Different stakeholder perspectives have been presented and analysed	All steakholder perspectives relevant to the intervention have been presented and analysed and how and why they have been selected is explained
	4c. Are conflicting findings and divergent perspectives presented and explained in the analysis and conclusions?	Divergent perspectives or conflicting findings are not presented	Diverging perspectives and conflicting findings are described	Divergent perspectives and conflicting findings are presented and explored	Divergent persepctives and conflicting findings are presented and explored, and there is an in-depth analysis of their implications for the conclusions

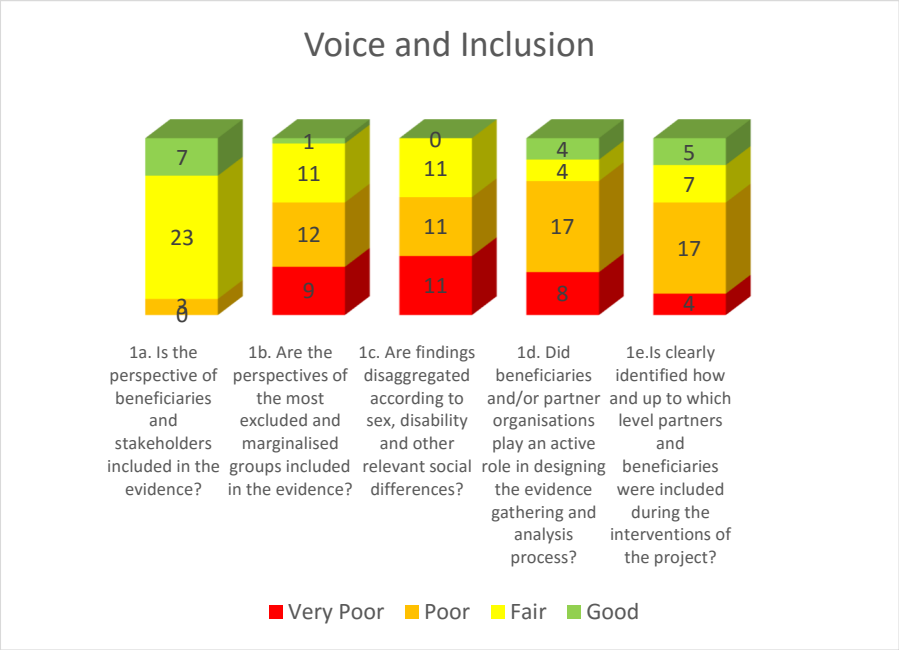
5) Contribution *of WV's interventions* We can show how change happened and explain how we contributed to it	5a. Is a point of comparison used to show that change has happened (e.g. baseline, a counterfactual, comparison with a similar group)?	No data is available to use as a point of comparison	Data is available and has been used as a point of comparsion	Data is available and has been used as a point of comparsion. A clear justification exists for why this is considered appropriate	Data is available and has been used as a point of comparison. A clear justification exists for why this is considered appropriate. The data provides a relevant and high quality basis for comparison.
	5b. Is the explanation of how the intervention contributes to change explored?	No causal link or assumptions are explored	Causal links between the intervention and outcomes are explored	Causal links between the intervention and outcomes and underlying assumptions are explored	All causal links between the intervention outcomes and underlying assumptions are explored in depth; the evidence provides a clear picture of whether the theory of underpinning the intervention's approach to change is sound
	5c. Are the alternative factors (e.g the contribution of other actors) explored to explain the observed results alongside our intervention´s contribution?	Analysis does not mentiion or explore the contribution of factors outside of the intervention	Analysis makes reference to the possible contribution of other factors outside of the intervention	Analysis explores and analyses the contribution of other factors outside the intervention	Analysis provides a comprehensive and systematic analysis of the relative contribution of other factors outside the intervention
	5d. Are unintended and unexpected changes (positive or negative) identified and explained?	Unintended changes are not explored	Unintended changes are identified	Unintended changes are identified and explained	Unintended changes are identified and explained. The methods used for data collection are designed to deliberately capture them.
6) Satisfaction of information needs	6a. Are the evaluation questions formulated in the report and clearly answered?	No evaluation questions are formulated	Evaluation questions are formulated, but not answered	Evaluation questions are formulated and to some extent answered	Evaluation questions are formulated and clearly answered

	6b. Are recommendations appropriate?	Recommendations are non-existent	Recommendations are general and unrealistic	recommendations are specific and useful, but do not assess viability and are not directed to specific actors	recommendations are project specified and their implementation is viable, they include considerations about resource constraints, directed to specific actors
	6c. Are lessons learnt highlighted?	Lessons learnt are non-existent	Lessons learnt are partially included in conclusions / recommendations	A few lessons learned are outlined, but no further explained	Lessons learned are outlined and explained
7) Conzeptualization of findings	7a. Were the findings structured along the Logframe objectives and indicators? (result table)	findings are not organized along the logframe indicators	some findings refer to logframe indicators	Most findings refer to logframe indicators	Findings are well structured and consistently organized along the logframe indicators
	7b.Is the report appropriate for different stakeholders in terms of scope and length, visualization, graphics and/or boxes? summarising table/charts baseline and actual data by indicators	The report is not appropriate for relevant stakeholders; summarising table is not available	The report is only appropriate for the donors due to ist complicated language, no visualization, no summaries, no overview of indicators	The report is appropriate for single stakeholders, facilitating understanding with some graphical elements.	the report is appropriate to all of ist relevant stakeholders in terms of ist scope, length, presentation and visualization of results. A summarising table is available.
	7c. Is the report structured well? E.g. following OECD / DAC criteria and/or specific evaluation questions?	there is no rationale for the used structure detected, organization of findings is rather confusing, e.g. DAC criteria are not considered at all	Some part of the report is well-structured along evaluation criteria	Most of the report is well-structured along evaluation criteria	general structure is logic, rationale can be followed easily, report is organized considering DAC criteria and/or specific evaluation questions

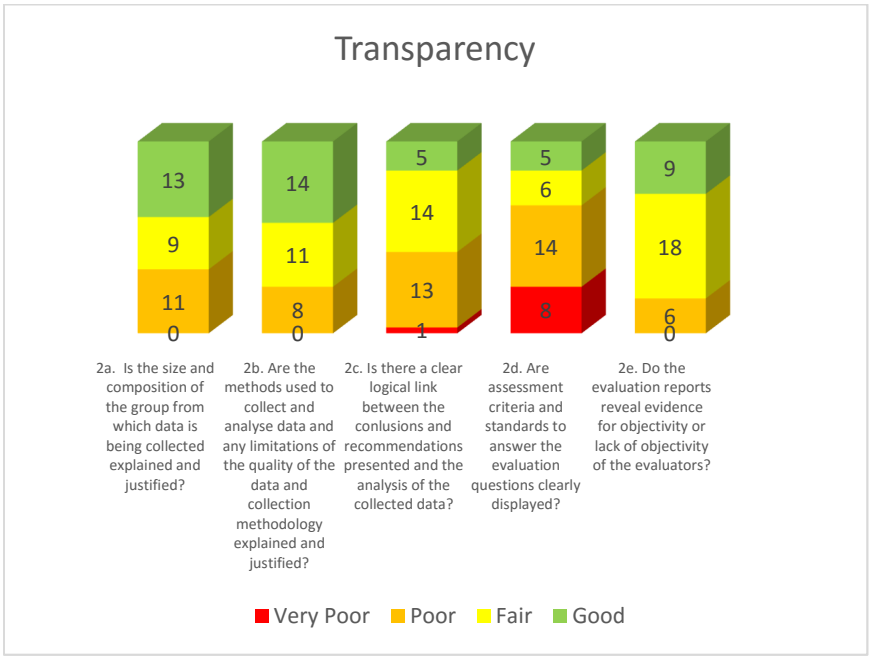
	7d. Does the final report include a well-structured executive summary covering all relevant aspects of the report?	The report does not contain an executive summary.	The report contains an executive summary, which covers some of the relevant aspects.	The report contains an executive summary, which covers most of the relevant aspects and is of satisfying structure.	The report contains an executive summary, which covers all of the relevant aspects and is very well-structured.
8) Sustainability	8a. Is the long term perspective of the interventions's sustainability included and any crucial aspects to consider analysed?	Sustainability of the intervention is not discussed	Sustainability of the intervention is not discussed sufficiently	Sustainability of the intervention is discussed covering only some crucial aspects	sustainability of the intervention is holistically discussed, covering all crucial aspects
	8b. Is the perspective of beneficaries on how to move on with the interventions after end of project included? (former 1e.)	The perspective of beneficiaries in terms of sustainability of the project is not considered at all	The perspective of beneficiaries in terms of sustainability is only briefly explained	Perspectives of beneficiaries are explained at one point in the report	Perspectives of beneficiaries are included in quali and quanti questionnaires, perspectives are presented and integrated into analysis
	8c. Are ownership and capacities of partner organization(s) to ensure sustainability of the intervention discussed?	Ownership and capacities of partner organizations are not discussed	Ownership or capacities of the partner organization are implicitly discussed	Ownership or capacities are discussed in the report to an extent without enabling the reader to understand the sustainability of the intervention	Ownership and capacities of partner organizations are discussed in depth and holistically, facilitating the understanding of sustainability of the intervention

6.2 Graphical illustration of sub-questions

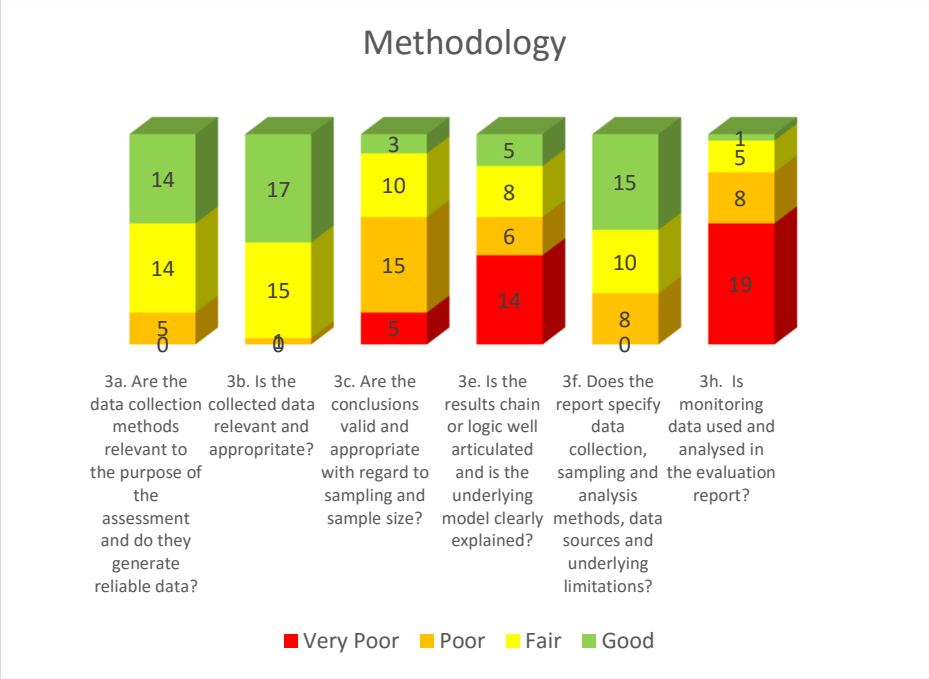
6.2.1 *Voice and Inclusion*



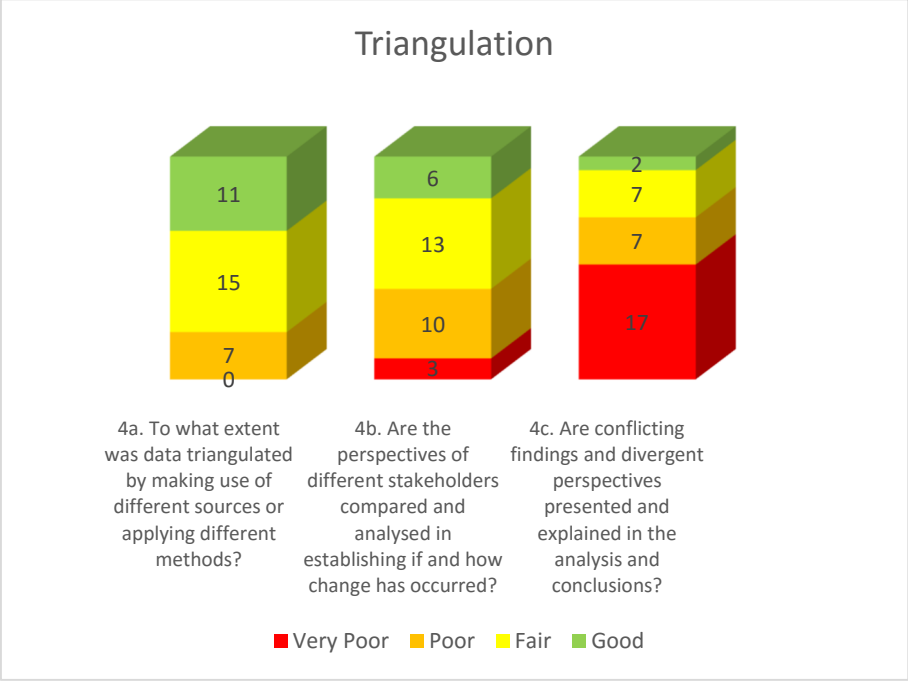
6.2.2 *Transparency*



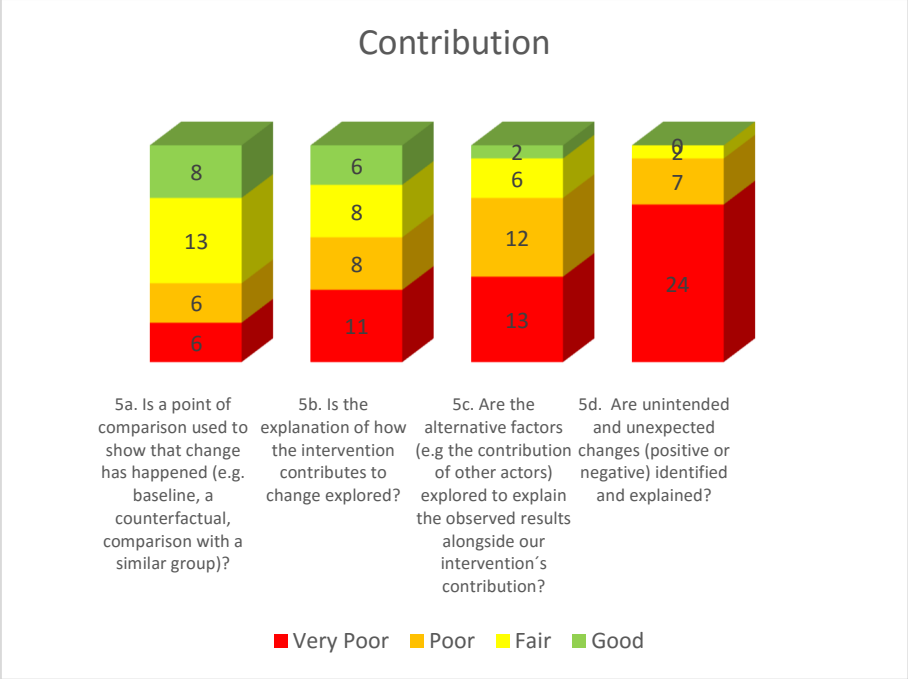
6.2.3 Methodology



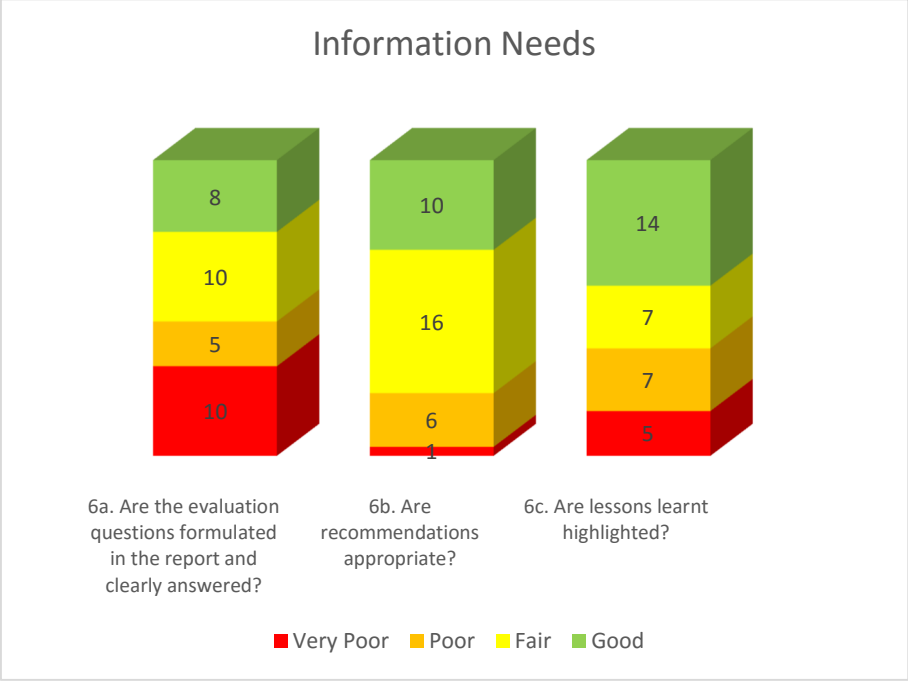
6.2.4 Triangulation



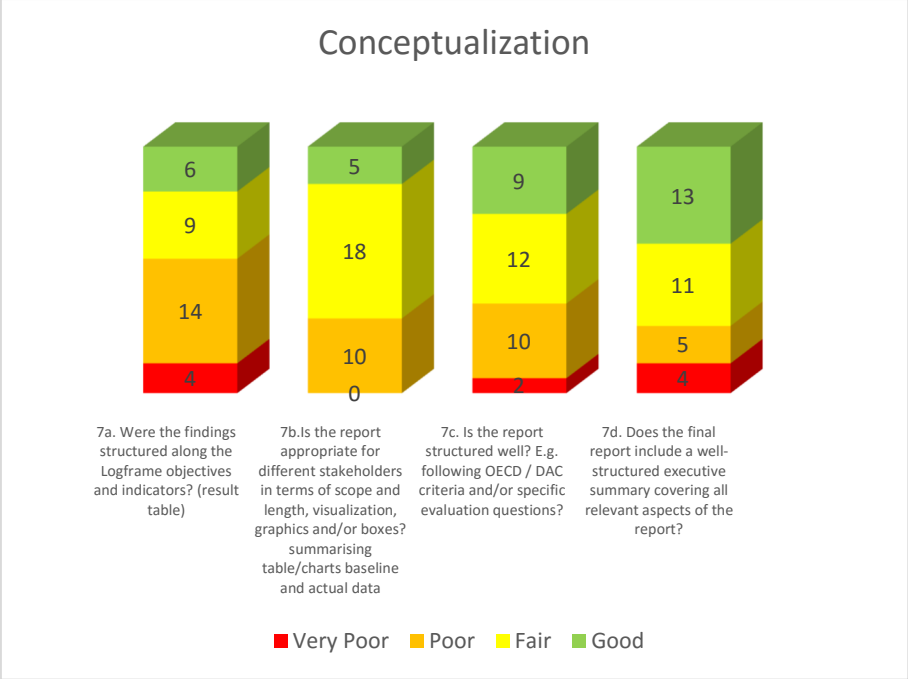
6.2.5 *Contribution of WV's interventions*



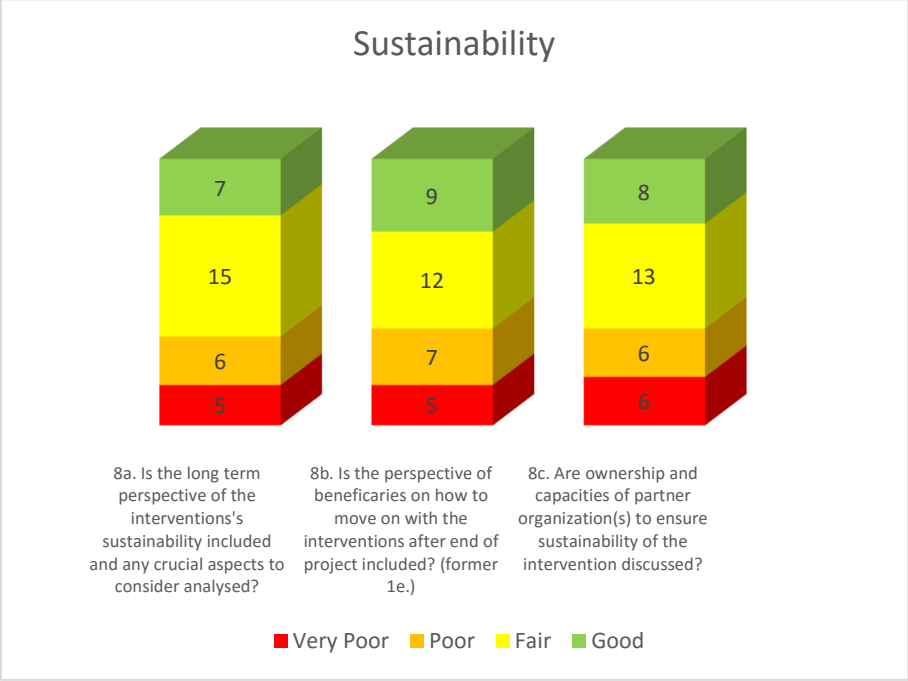
6.2.6 *Satisfaction of information needs*



6.2.7 Conceptualization of Findings



6.2.8 Sustainability



6.3 Summary of Results