

# Meta-Evaluation of 29 evaluation reports of World Vision Germany

on behalf of  
World Vision Germany

**Authors:**

Sophie Wannemacher  
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  
s.wannemacher@ceval.de  
URL <http://www.ceval.de>

Saarbrücken, August 2016

**Table of contents**

1. Executive Summary in German (Kurzzusammenfassung) ..... 1

2. Background..... 6

3. Methodological Approach..... 7

4. Assessment according to quality of evaluation criteria ..... 8

    4.1 Voice and inclusion of beneficiaries ..... 8

    4.2 Transparency ..... 10

    4.3 Appropriateness of evaluation methods..... 11

    4.4 Methodology ..... 12

    4.5 Triangulation ..... 14

    4.6 Identification of WV’s contribution..... 15

    4.7 Satisfaction of information needs ..... 17

    4.8 Organization of findings ..... 18

5. Further findings from the online survey..... 19

6. Conclusion ..... 21

7. Recommendations..... 24

8. Annex..... 26

    8.1 Criteria and Sub-criteria ..... 26

    8.2 Graphical Illustration of Sub-criteria ..... 26

        8.2.1 Voice and Inclusion..... 26

        8.2.2 Transparency ..... 27

        8.2.3 Appropriateness ..... 28

        8.2.4 Methodology ..... 29

        8.2.5 Triangulation ..... 30

        8.2.6 Contribution of WV’s interventions ..... 30

        8.2.7 Satisfaction of information needs ..... 31

    8.3 Online-Survey Questionnaire ..... 31

    8.4 Online-Survey Results..... 31

## 1. Executive Summary in German (Kurzzusammenfassung)

Im Jahr 2014 hat World Vision Deutschland (WVD) das Centrum für Evaluation (CEval) erstmalig mit der Umsetzung einer Meta-Evaluation beauftragt. Ziel dieser Meta-Evaluation war es auf Grundlage der Untersuchungsdimensionen Mitsprache und Inklusion, Transparenz, Angemessenheit der Methoden, Methodik, Triangulation, Identifizierung des Programmbeitrags sowie Befriedigung des Informationsbedarfs von WV die Qualität der einzelnen Berichte zu bewerten, um anschließend eine Bewertung der Wirkungsberichterstattung an WV vorzunehmen. Die zentralen Befunde der Meta-Evaluation wurden von WVD im letzten Wirkungsbericht 2014<sup>1</sup> veröffentlicht.

Auch dieses Jahr wurde die Qualität von 29 Evaluationsberichten aus den Finanzjahren 2014 und 2015 durch das CEval bewertet. Hierbei wurden die o.g., in der vorangegangenen Studie angewandten Untersuchungsdimensionen erneut zugrunde gelegt, jedoch um das Kriterium „Ergebnisdarstellung innerhalb des Berichts“ ergänzt. Dieses Kriterium bewertet den Aufbau des Evaluationsberichts. Ist dieser beispielsweise nach den DAC Kriterien und/oder der Planungsmatrix aufgebaut? Liegt die Analyse einem Wirkungsgefüge (Theory of Change) zu Grunde und ist die von WV zu Anfang gewünschte Übersichtstabelle mit den wichtigsten quantitativen Ergebnissen vorhanden?

Nach dem gleichen Prinzip wie 2014 wurde ein zweistufiges Auswertungsverfahren angewandt. In einem ersten Schritt wurden die Kriterien anhand von verschiedenen Unterkriterien mittels einer vierstufigen Skala (von 1 = mit erheblichem Verbesserungsbedarf bis 4 = gut bis sehr gut) bewertet. Da alle Unterkriterien gleichmäßig gewichtet wurden, erfolgte die im zweiten Schritt vorgenommene aggregierte Bewertung durch die Berechnung des Mittelwerts.

Zusätzlich wurde ein Überblick über die Anwendung innovativer qualitativer Instrumente und die Anwendung spezieller quantitativer Instrumente erstellt<sup>2</sup>. Außerdem wurde erhoben, bei wie vielen Evaluationen eine Vergleichsgruppe mit einbezogen und bei wie vielen Auswertungen Baseline Daten verwendet wurden.

Der Analyserahmen wurde bei der diesjährigen Meta-Evaluation um Nutzung und Nützlichkeit von Evaluationsergebnissen für die Projektarbeit erweitert. Die Nützlichkeitsbewertung erfolgte dabei auf Grundlage einer internationalen, organisationsweiten Online-Befragung der World Vision Mitarbeiter.

130 Mitarbeiter weltweit wurden dazu aufgefordert an der Befragung teilzunehmen. Dazu wurden Teilnehmer auf verschiedenen Organisationsebenen in den Projekt-, Länder-, Unterstützer- (wie z. B. World Vision Deutschland) sowie Regionalbüros,<sup>3</sup> angefragt. Laut WVD werden durch die o.g. Anzahl die relevanten Mitarbeiter im Umgang mit Evaluationen gut abgebildet. Der teilstandardisierte Fragebogen enthielt Fragen zur Qualität der intern vorliegenden Evaluationen, um die Ergebnisse aus der Dokumentenanalyse mit der Qualitätseinschätzung der Mitarbeiter zu vergleichen. Daher finden sich im Online Fragebogen teilweise die Untersuchungskriterien, die auch der Dokumentenanalyse zu Grunde liegen wieder. Zusätzlich wurden gezielt insbesondere qualitative Daten zu Verwendung, Verbreitung, Nutzen und Nützlichkeit der Evaluationsergebnisse erhoben.

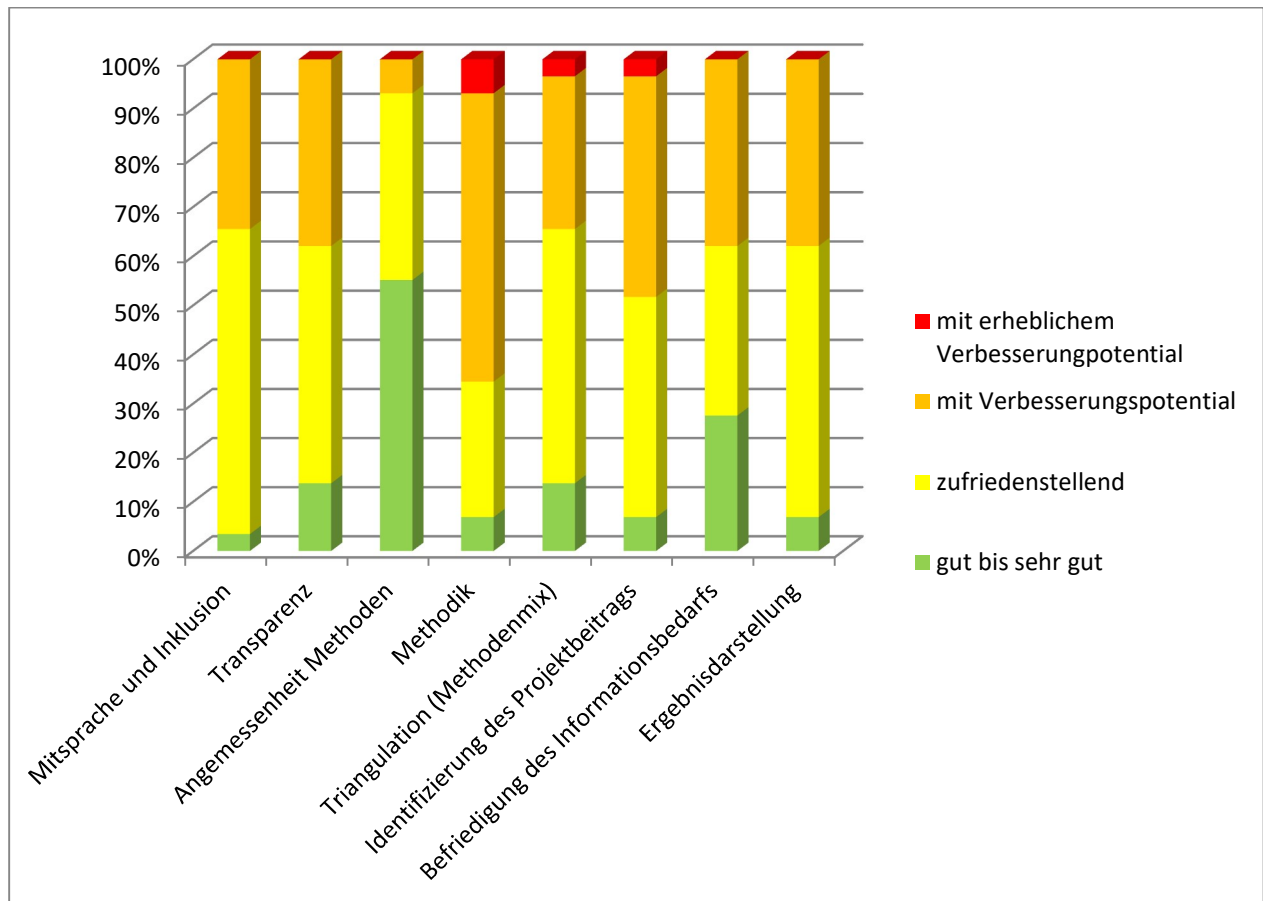
<sup>1</sup> [https://www.worldvision.de/\\_downloads/allgemein/14wbde.pdf](https://www.worldvision.de/_downloads/allgemein/14wbde.pdf) (Stand 19.05.2016)

<sup>2</sup> Detaillierte Informationen zu diesen, später im Text genannten Instrumenten sind online verfügbar: <http://www.wvi.org/development/publication/functional-literacy-assessment-tool-flat> (Stand 23.05.2016)

<sup>3</sup> Siehe Jahresbericht 2015, S.18 und 19 für einen Überblick über die Organisationsstruktur

Die Ergebnisse der diesjährigen Meta-Evaluation sind denen der Bewertung vor zwei Jahren sehr ähnlich. Abbildung 1 zeigt, dass wie damals, die Mehrheit der Evaluationsberichte einen Großteil der untersuchten Kriterien mindestens zufriedenstellend erfüllt. Angemessenheit der ausgewählten Methoden (16 Berichte gut bis sehr gut, 11 zufriedenstellend) und Befriedigung des Informationsbedarfs (8 Berichte gut bis sehr gut, 10 zufriedenstellend) sind erneut die am besten bewerteten Kriterien. Weiterhin weisen immer noch zahlreiche Evaluationen Verbesserungspotential im Hinblick auf Mitsprache und Inklusion, Transparenz, Methodik, Triangulation, Identifizierung des Projektbeitrags und Ergebnisdarstellung auf.

Abbildung 1: Zusammenfassung der Ergebnisse



Auffällig ist der Fortschritt bei der Identifizierung des Programmbeitrages. Dies ist darauf zurückzuführen, dass deutlich öfter versucht wurde, einen Vergleich zum Zeitpunkt vor Programmbeginn herzustellen. Während nur bei der Hälfte der Evaluationsberichte aus den Finanzjahren 2012 und 2013 Baseline Daten genutzt wurden, wurde bei Evaluationen im aktuellen Untersuchungszeitraum bei gut 80% der Vergleich zum Zeitpunkt vor Programmbeginn hergestellt. Außerdem wurde mehr Wert darauf gelegt, positive und gegebenenfalls negative nicht intendierte Effekte zu identifizieren und zu erklären. Auch auf die Einflüsse externer Faktoren, beispielsweise staatliche Entwicklungsprogramme, Aktivitäten anderer Geber oder Umweltveränderungen wurde verstärkt hingewiesen. Aufgrund der meist unzureichenden Verfügbarkeit oder Qualität von Baseline Daten und dem nach wie vor meist fehlenden Vergleich mit einer adäquaten Vergleichsgruppe konnten die von WV verursachten positiven oder negativen Effekte jedoch statistisch nicht isoliert, und somit der Programmbeitrag nicht eindeutig identifiziert werden.

Deshalb sollte neben dem Einbezug einer Kontrollgruppe, insbesondere auch auf die Erhebung valider Baseline Daten geachtet werden. Zusätzlich gilt, die Vergleichbarkeit der Indikatoren bei der Folgerhebung sicherzustellen. Letztlich sollten idealerweise die gleichen Haushalte befragt (Panelbefragung) werden.

Auch die Darstellung und Erklärung von Ursache-Wirkungszusammenhängen würde helfen, nachvollziehen zu können, wie die jeweiligen Interventionen zu den identifizierten Veränderungen beitragen konnten. In den meisten Berichten fehlt jedoch die Darstellung einer solchen Wirkungslogik. Damit das Bewertungsergebnis für jeden Leser nachvollziehbar ist, empfiehlt das CEval sicherzustellen, dass diese Zusammenhänge im Bericht kurz, eventuell graphisch, dargestellt werden.

Während zur Erhebung qualitativer Daten meist auf die klassischen Methoden der Fokusgruppendifkussion sowie Experteninterviews zurückgegriffen wird, stieg die Nutzung spezifizierter quantitativer Datenerhebungsinstrumente deutlich an. Die nachfolgende Tabelle zeigt, dass insbesondere der Test zur Einstufung der Kinder als „funktional lesefähig“ bei Evaluationen in den letzten beiden Finanzjahren weitaus häufiger genutzt wurde als in den Finanzjahren 2012 und 2013. Auch der Anteil der Evaluationen die zur Datenerhebung das Caregiver Survey Instrument und/oder das Youth Health Behaviour Survey (YHBS) Instrument nutzten, ist deutlich angestiegen. Das Caregiver Survey Instrument fand im aktuellen Untersuchungszeitraum bei knapp der Hälfte der Evaluationen Anwendung. Auch die Nutzung der YHBS hat sich von 15 auf 38 Prozent deutlich erhöht.

**Tabelle 1**

Nutzung von WV's Datenerhebungsinstrumenten	Finanzjahre 2012 & 2013 N=34		Finanzjahre 2014 & 2015 N=29	
	N	In%	N	In%
Functional Literacy Assessment Tool (FLAT)	5	15%	21	72%
Development Asset Profile (DAP)	3	8%	7	24%
Caregiver Survey	6	18%	14	48%
Youth Health Behaviour Survey (YHBS)	5	15%	11	38%
Measuring child growth	8	24%	9	31%
<b>Anwendung Innovativer Qualitativer Methoden</b>				
Comparison discussion group	4	12%	0	0%
Photo-voice	2	6%	1	3%
Seed assessment	5	15%	2	7%
Ladder of life	3	8%	4	14%
Tree of change	7	21%	7	24%

Nachdem ein kurzer Überblick über die unabhängige Bewertung der Evaluationsberichte und die Nutzung der Erhebungsinstrumente gegeben wurde, werden nachfolgend die Befunde aus der Berichtsanalyse mit den Einschätzungen der Mitarbeiter von WV verglichen und die zusätzlichen Ergebnisse aus der Online-Befragung erläutert.

Der Rücklauf der Online-Befragung betrug circa 50%. Die Gruppe der Länderbüros ist dabei mit knapp mehr als der Hälfte der Teilnehmer am besten vertreten.

Die Ergebnisse der Befragung zeigen, dass die Mitarbeiter von WV die Kriterien Transparenz, Methodik, Identifizierung des Programmbeitrags, Befriedigung des Informationsbedarfs und Ergebnisdarstellung tendenziell besser bewerten als das Expertenteam. Hierbei ist jedoch zu berücksichtigen, dass sowohl die Bewertungsgrundlage als auch der Maßstab ein jeweils anderer war. Während die Mitarbeiter unterschiedliche Berichte vor unterschiedlichen Erfahrungshintergründen und mit verschiedenen Erwartungshaltungen bewerteten, fällte das Expertenteam sein Urteil auf Grundlage von 29 an WV Deutschland gerichtete Berichte und mit Hilfe eines festgelegten Kriterienrasters.

Dennoch bestätigt sich die in der Dokumentenanalyse festgestellte, immer noch vorhandene Schwachstelle bei der Identifizierung des Programmbeitrags. Bei der Frage nach der Qualität der Evaluationen hinsichtlich der oben genannten Kriterien ist dieses Kriterium das am schlechtesten bewertete. Während alle anderen Kriterien nur von unter 5% der Teilnehmer (Ausnahme Methodologie 6,3%) „mit erheblichem Verbesserungsbedarf“ bewertet werden, sehen 11% der Teilnehmer Verbesserungsbedarf bei der Identifizierung des Programmbeitrags, 14,3% sogar erheblichen Verbesserungsbedarf. Für gut 20% der Teilnehmer ist die Identifizierung des Programmbeitrags lediglich zufriedenstellend. Wie ebenfalls in der Dokumentenanalyse festgestellt, zeigt die qualitative Auswertung der Online-Befragungsergebnisse, dass dies in den Augen der Mitarbeiter meistens der niedrigen Qualität erhobener Baseline-Daten oder gar der mangelnden Verfügbarkeit solcher Daten geschuldet ist. Alle Befragten, die die Bewertung dieses Kriteriums kommentierten, nennen ebenfalls das Fehlen einer adäquaten Kontrollgruppe als Begründung für die unzureichende Identifizierung des Programmbeitrags.

Dieser Befund steht im Einklang mit dem Antwortverhalten der Teilnehmer bei der Frage nach den wichtigsten Aspekten für die Qualität von Evaluationen. Für 42,6% der Teilnehmer steht das Evaluationsdesign an erster Stelle. Nur die Qualifikation des Evaluationsteams wird noch öfter als wichtigstes Kriterium genannt (von 45,1%).

Bei der Frage nach anderen möglichen Einflussfaktoren auf die Qualität von Evaluationsberichten, wird neben der Professionalität des Evaluationsteams ähnlich oft die Einbeziehung der Interventionsgemeinde bzw. der Begünstigten in den Evaluationsprozess genannt. Obwohl die Nützlichkeit der Empfehlungen bereits von zwei Dritteln der Befragten als mindestens gut bewertet wurde, erhofft man sich durch die Diskussion der Ergebnisse mit den Zielgruppen insbesondere spezifischere Empfehlungen ableiten zu können, zudem aber auch die Nachhaltigkeit der Projekte durch die Kommunikation von motivierenden, positiven Ergebnissen zu sichern.

Was die generelle Qualität der Evaluationen betrifft, stellten 57,4% der Befragten in den vergangenen zwei Jahren eine leichte Verbesserung, 26,2% sogar eine erhebliche Verbesserung fest. Der Anteil der Mitarbeiter, die eine Verschlechterung in der Qualität der Berichte in diesem Zeitraum wahrgenommen hat, liegt bei unter 5%. Der Trend hin zur Verbesserung deckt sich ebenfalls mit den Ergebnissen der Dokumentenanalyse.

Hinsichtlich des konkreten Nutzens der Evaluationen sind sich alle Teilnehmer einig, dass dieser hauptsächlich darin besteht, Projekte zukünftig besser gestalten zu können. D.h. sowohl laufende Projekte, bei denen die gewonnenen Erkenntnisse direkt nach einer Zwischenevaluation genutzt werden können um Änderungen im Projektdesign vorzunehmen, als auch neue Projekte, bei deren Design die Ergebnisse aus Zwischen- und Endevaluationen anderer Projekte mit berücksichtigt werden können. Da in letzterem Fall jedoch Kontext und Stakeholder variieren und eine direkte Anwendung der Erkenntnisse somit nicht möglich ist, wäre zu erwarten, dass Zwischenevaluationen nützlicher bewertet werden, als Endevaluationen. Dies ist jedoch nicht der Fall. Die Anteile der Befragten, die Zwischenevaluationen bzw. Endevaluationen als jeweils nützlicher erachten halten sich mit jeweils 17,7% die Waage, während deutlich mehr als die Hälfte der Befragten (56,5%) diesbezüglich keine Präferenz äußert.

Auch das Einfließen der Ergebnisse in die Child Well Being Reports<sup>4</sup> wird vom Großteil der Befragten bei der Frage nach der Nutzung der Evaluationsergebnisse genannt. Außerdem müssen die Evaluationsergebnisse den Unterstützungsbüros und den Gebern präsentiert werden. Die qualitative Datenanalyse zeigt aber auch das Vorhandensein eines Bewusstseins für die Bedeutung der Verbreitung der Ergebnisse in Regierungsinstitutionen und anderen Nicht-Regierungsinstitutionen um Anreize für ähnliche Projekte zu schaffen oder Erfahrungswerte weiterzugeben und somit bereits bestehende oder neue Projekte dieser Institutionen positiv beeinflussen zu können. Bei der Frage nach der Verbreitung von Evaluationsergebnissen wird ebenfalls erneut intensiv auf die Wichtigkeit der Kommunikation der Ergebnisse mit den Bewohnern der Interventionsgemeinden hingewiesen, um diese in der Weiterführung der verschiedenen durchgeführten Projekte zu bestärken.

Die diesjährige Meta-Evaluation identifiziert die gleichen Defizite wie die vorherige Meta-Evaluation. Nichtsdestotrotz zeigt sich, dass sich insbesondere die Anwendung innovativer quantitativer Methoden und die Identifizierung des Programmbeitrags verbessert haben. Die Online Befragung bestätigt die Schwachstellen aber auch den insgesamt positiven Trend. Aus ihr geht außerdem hervor, dass die Evaluationen von den Mitarbeitern generell, ob extern, intern, Zwischen- oder Endevaluationen mehrheitlich als nützlich bewertet werden. Die Ergebnisse werden in den Lernprozess mit einbezogen und finden somit letztlich meist in der zukünftigen Projektgestaltung Verwendung. Darüber hinaus fließen die gesammelten Erfahrungen auch häufig in die Projektplanung anderer Institutionen mit ein. Wünschenswert wäre neben mehr quasi-experimentellen Ansätzen auch die vermehrte Einbeziehung der Zielgruppen in den Evaluationsprozess. Dies hätte nicht nur den von den WV Mitarbeitern beschriebenen Vorteil für den Nutzen der abgeleiteten Empfehlungen, sondern würde darüber hinaus die Akzeptanz der gegebenen Empfehlungen bei der lokalen Bevölkerung erhöhen, sodass von einem positiven Einfluss auf die Weiterführung der Projekte auszugehen ist.

---

<sup>4</sup> Diese Reports werden seit 2011 von World Vision veröffentlicht und bieten einen Überblick der erreichten Ziele an Hand der eigens definierten Child Well Being Outcomes ([http://www.wvi.org/sites/default/files/Child\\_Well-being\\_Aspirations\\_and\\_Outcomes\\_English\\_0.pdf](http://www.wvi.org/sites/default/files/Child_Well-being_Aspirations_and_Outcomes_English_0.pdf))

## 2. Background

To maintain its programmes and activities a non-governmental development organization like World Vision Germany (WVG) relies on a continuous inflow of private funds. In times of a vast supply on initiatives and a vibrant NGO scene, a philanthropic approach seems to be no more sufficient to ensure the success of fundraising campaigns. Since private donors are sensitised to infidelity and misappropriation, non-governmental organisations are much more required to prove relevance, effectiveness, efficiency, impact and sustainability of their programmes than in earlier decades. Thus, evaluation is a key tool to improve performance and accountability of an organisation and its interventions.

Like in 2014, WVG engaged the Centre for Evaluation (CEval) to conduct a meta-evaluation in the thematic areas of child well-being and community development to assess the methodological quality of its evaluation reports. CEval thereby reviewed several criteria<sup>5</sup> developed by WVG (with impulses of CEval) to evaluate the quality of evaluation reports presented to WVG. The assessment is structured according to the criteria:

1. Voice and inclusion of beneficiaries,
2. Transparency,
3. Appropriateness of evaluation methods,
4. Methodology,
5. Triangulation,
6. Identification of WV's contribution,
7. Satisfaction of information needs
8. Organization of findings

This year, additionally an online survey investigating the utility, usage and usefulness of evaluation findings within WV has been conducted. 130 WV staff members have been invited to participate in the online survey. Staff members from different organizational levels (project offices, national offices, support offices and district/base offices) and various countries have been contacted by WVG. According to WVG the above mentioned number adequately depicts the relevant employees dealing with evaluations. The semi-standardized questionnaire contained questions concerning the quality of evaluation reports against several criteria, similar to the above mentioned ones, to compare the perception of quality within WV with the results from the document analysis. Furthermore, qualitative data has been collected to get insights into the usage, utility and usefulness of evaluation results.

The remainder of this report starts with a brief introduction of the methodological approach to this meta-evaluation (Chapter 3). Furthermore, a stepwise analysis of every single criterion will be presented and if possible compared or connected to the results of the online survey (Chapter 4) before additional findings from the online survey will be outlined (Chapter 5). After a brief conclusion (Chapter 6), this report offers recommendations (Chapter 7) to improve future evaluation quality.

---

<sup>5</sup> For an overview of the criteria including the respective subcriteria see the Excel sheet in Annex 8.1



### 3. Methodological Approach

This meta-evaluation is based on 29 reports which have been produced between October 2013 and September 2015. They summarise the evaluation process of WVG's Area Development Programmes (ADP) in Sub-Sahara-African, Asian and Latin American countries, and are thus written in English (25), French (2) or Spanish (2). To the best of CEval's knowledge this is a complete sample of all ADP evaluations conducted in this time period.

An ADP can be understood as a programme 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. Thus, this meta-evaluation is based on both, mid-term and final evaluations. Although many reports draw on baseline data, such data is not always available. Hence, as ADPs vary, so did the preconditions for the evaluation teams. Accordingly, this meta-evaluation cannot offer a sound comparison of the evaluation reports with each other, but rather assesses their individual quality against the eight mentioned criteria.

In doing so, a checklist serves as an analysis grid and a grading system which differentiates between the four categories: very poor, poor, fair, and good quality will be implemented. To avoid oversimplification and to allow a wide range of different aspects, the analysis starts with the assessment of three to seven sub-criteria for each of the eight criteria (i.e. voice and inclusion, transparency, appropriateness of evaluation methods, methodology, triangulation, contribution, satisfaction of information needs and organization of findings). Furthermore, findings will be consolidated and one aggregated rating – the mean – per criterion will be deviated.

The scale definitions provided by the BOND tool<sup>6</sup> allow the classification into one of the four categories. For the additional criteria and thus sub-criteria added by CEval, scale definitions have been developed following the logic of the BOND tool. Due to the rather extensive size of the reports<sup>7</sup>, limited resources, and hence, time constraints, the scale definitions could only serve as an orientation. Reports could not be assessed in detail. This holds especially true for analysing the content in an all-embracing manner. Therefore, the focus lies on methodological issues, and on providing evidence on de facto application of WV's data collection instruments, application of innovative qualitative methods and appropriateness of interpreting quantitative data to provide a hint on challenges and to highlight promising evaluation cases. Furthermore, for each report, specific sections are assessed to inquire how well methods were applied and how valid findings were deviated. By doing so, this meta-evaluation identifies general trends, displays heterogeneity, and prepares the ground for enhancing the quality of ADP evaluation.

A comparison of the overall result of this year's document analysis and the one from two years ago only restrictively allows conclusions since the changes in budget, conditions and guidelines within the timeframe could not have been analysed within the scope of this study.

---

<sup>6</sup> Tool used by WV to assess the quality of evidence: <https://www.bond.org.uk/effectiveness/principles#download>

<sup>7</sup> Including the terms of reference which have to be considered to evaluate the "satisfaction of information needs" criterion some reports count over 100 pages.

For the development of the online survey three semi-structured interviews concerning the quality of evaluation reports, the utility, usage and usefulness of evaluation reports within WV have been conducted (via Skype) with staff from different office types in different countries. Based on the findings and impulses from these interviews the online survey questionnaire has been developed.

To find out if the results from the document analysis are congruent with the perception of the quality of evaluation reports within WV's staff, also the checklist for the document analysis served as a base for the online questionnaire. Nevertheless, aiming for a high number of participants, the questionnaire was kept as short as possible, which is why it only partially contains the earlier mentioned criteria. Furthermore, it contained additional questions concerning quality, the development of quality over time, utility, usage and usefulness of evaluation findings within WV.

The online tool contained a mixture of standardized and non-standardized questions. 130 staff members from different organizational levels (project offices, national offices, support offices and district/base offices) and various countries, handling evaluations have been invited by WVG to fill out the short questionnaire. The net response rate was about 50%, which is why the sample size shrunk to 63 participants. As 35 of the participants work in the national offices, a differentiation between the office levels would not be representative<sup>8</sup>. District/base office and project office each represent almost 15% of the participants. Staff of the support offices is the smallest group with round about 11%.

The results of the online survey are used to assess if the internal perception of quality is congruent with the findings from the document analysis. Nevertheless, only tendencies can be compared. A direct comparison would not be valid since each survey participant took his or her decision on a different data base, with different backgrounds and hence different expectations while the expert team's assessment was made against defined criteria and 29 specific evaluation reports presented to WVG.

## 4. Assessment according to quality of evaluation criteria

Hereinafter, the aggregated results of all criteria and sub-criteria are discussed and the aggregated result for each criterion is graphically outlined. A graphical representation of the aggregated results of each sub criterion can be found in Annex Chapter 8.2.

### 4.1 Voice and inclusion of beneficiaries

This section highlights to which extent the perspectives of the beneficiaries and stakeholders are included in the evaluation reports. It further shows whether the views of the most excluded and marginalised groups are adequately incorporated and whether findings were appropriately disaggregated according to sex, disability or other causes of social differentiation. Nearly all evaluation reports (25 out of 29) perform good or fair with regard to capturing the voice of a wide range of stakeholders. In that regard especially the report of the Mukumbura ADP in Zimbabwe can be highlighted

---

<sup>8</sup>  $N_{support\ office} = 7, N_{District/Base} = 9, N_{Project} = 9$ . 3 participants did not choose an affiliation. It is not clear if either their kind of office was not included or if they left that information out on purpose, nevertheless they accurately filled the rest of the questionnaire, which is why the data has also been included in the analysis.

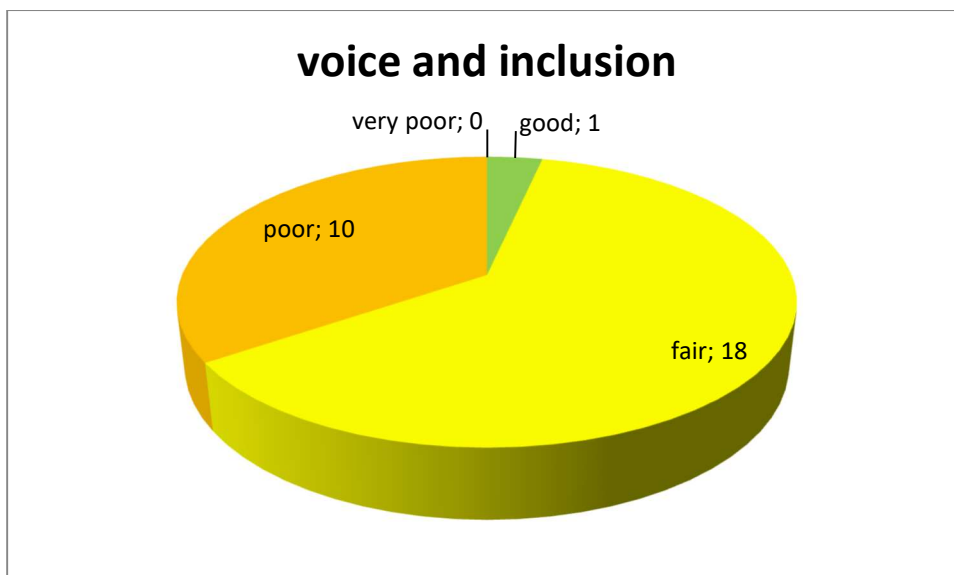
as a good example for future evaluations as it does use various data collection instruments (caregiver survey, youth health behaviour survey, tree of change, key informant interviews and various focus group discussions with different beneficiary groups) to capture the voice of diverse stakeholders appropriately. In contrast, only about half of the reports present disaggregated findings appropriately (8 rated good, 7 fair, 14 poor, 0 very poor).

Our analysis furthermore reveals that the beneficiaries do not play an active role in designing the evidence gathering and analysis process. Only 5 reports perform fair in this sub criterion while the rest performs poor (4) or very poor (20). In a few reports it is stated that a participatory approach has been implemented. Nevertheless, it is not further explained how beneficiaries and other stakeholders were involved in the design phase of an evaluation (as specified by the sub-criterion). However, to be fair on assessment, WVG may rethink to which extent participation of beneficiaries and other stakeholders will be de facto feasible at this stage, given inherent budget and time constraints of its evaluations.

In contrast, it is state-of-the-art to evaluate how and up to which level beneficiaries and other stakeholders were included during the interventions of a project. Hence, it is surprising that 17 reports are either rated poor or very poor, 10 perform fair and only 2 out of 29 reports present clear evidence on this issue.

Although it is of major importance for the sustainability of the project, the inclusion of beneficiaries' perspectives on how to move on with the interventions after the end of WV's programmes is often not addressed appropriately (9 good, 5 fair, 11 poor, 4 very poor). Nevertheless, observations made and impressions gained by the evaluation team concerning the long-term perspective of the interventions are presented in most of the reports (14 are rated well, 8 fair, 3 poor and 4 very poor).

Figure 1: Overall performance referring to the criterion voice and inclusion



To sum up, **Fehler! Verweisquelle konnte nicht gefunden werden.** presents aggregated findings and reveals that overall one report performs well, while 18 reports perform fair, and 10 are rather poor with regard to WVG's evaluation criterion "voice and inclusion".

The results from the online survey confirm the findings from the document analysis. Also the participants of the online survey opt for more participatory approaches. The wish to further include the local community especially into the feedback process is often mentioned by WV staff members to improve the quality and utility of evaluation results. Overall quality is stated to be affected positively through the participation of beneficiaries because the contribution of the local community enables the evaluation team to derive and formulate more realistic, precise and hence, more useful recommendations. The positive effect on utility of the evaluation results is seen when sharing the results with the community because positive results are assumed to nourish beneficiaries' further motivation and participation and hence, sustainability of the projects can be fostered.

## 4.2 Transparency

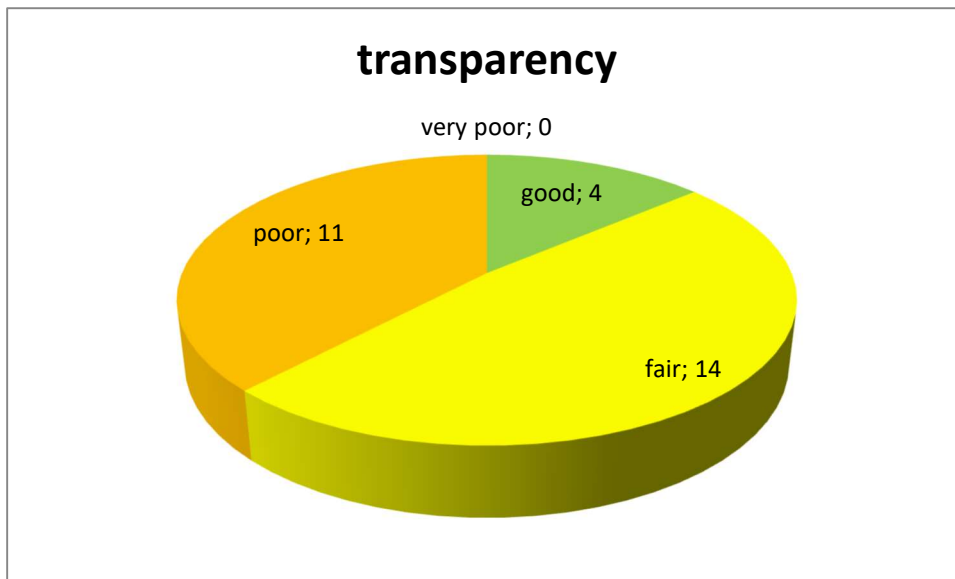
The transparency of an evaluation is characterised by openness about data sources, methods, limitations and the affiliation of evaluators. This section subsumes these aspects under WVG's transparency criterion. Detailed assessment reveals that several evaluation reports lack a proper discussion on the limitations of both collected data and applied methodology (13 rated poor, 2 very poor). Only 5 reports describe and justify the size and the composition of their data sets well. Furthermore, 9 reports are rated fair in disclosing the rationale behind their empirical base. Nevertheless, 8 reports clearly present a need for improvement in terms of reliability or absence of data sources and valid conclusions.

An additional weakness is that some reports do not appropriately disclose information about its authors and their institutional affiliations. Furthermore, most of the evaluations were implemented by WV's staff. Beyond biases through the author, also interviewers can bias the results. Thus, it is a further point of critique that data for some evaluation reports was collected by WV staff.

Another challenge of the reports is to establish a logical link between the analysis conducted and their recommendations provided. Performance is acceptable concerning this sub-criterion: 12 reports perform well and 6 fair. Nevertheless, there is room for improvement since 9 reports hardly establish this link and 2 even completely fail to do so.

**Fehler! Verweisquelle konnte nicht gefunden werden.** summarises that overall round about 60% of the evaluation reports comply at least fairly with WVG's transparency criterion.

Figure 2: Overall performance referring to the criterion transparency



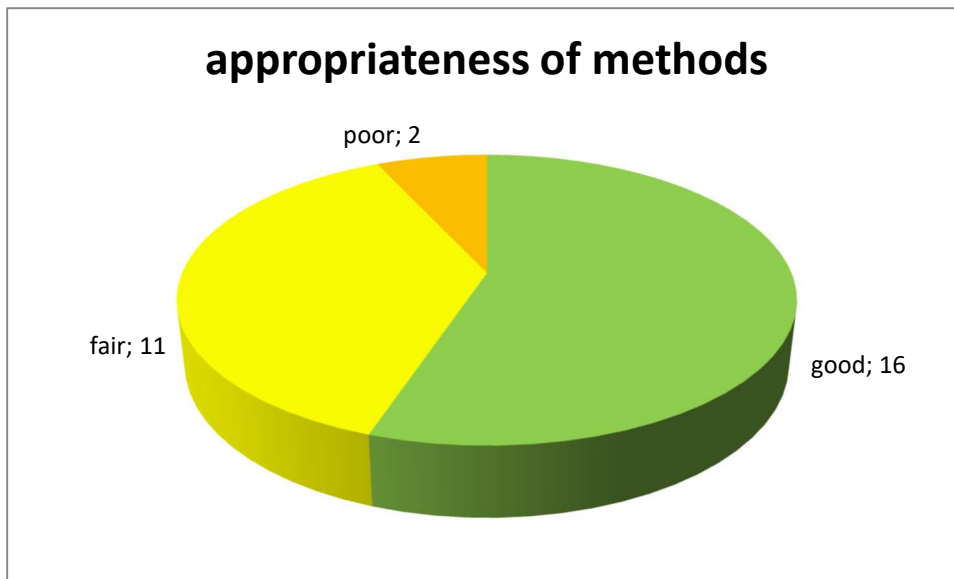
While nearly half of the reports (14) perform fairly, 4 evaluations fulfil the criterion fully. Nonetheless, as 11 reports do not meet WVG's quality standards in a sufficient way, CEval suggests taking the evaluation report of the Vardenis ADP in Armenia as promising example to address relevant aspects of transparency in a good manner by comprehensively disclosing all steps of the analysis including underlying data sources, data collection strategies and quality assurance.

The results of the online survey are in line with the overall result of the transparency criterion. Amongst a series of other aspects, LEAP contains a detailed guideline how to report in a transparent manner. Its helpfulness is often mentioned among the staff members and most of them state to easily comply with these guidelines which is why a certain confidence dominates among participants. 60% evaluate the transparency at least "good". Nearly 20% even "very good". Only 11% assesses it as fair while the percentage which states any need for improvement is below 10%.

### 4.3 Appropriateness of evaluation methods

Whereas the methodology criterion (Chapter 4.4) focuses on the articulation of the results chain, the quality of the used data, and analysis methods, this section reveals whether applied methods are appropriate given the nature of the intervention, the purpose of the assessment and the evaluation framework of WVG. The analysis reveals that 24 of the 29 reports state data collection methods, which are relevant to the purpose of ADP evaluation (4 rated fair, 1 poor). This positive picture also holds for the appropriateness of collected data (16 rated good, 11 fair, 2 poor). This yields to a quite positive rating of overall appropriateness of methods (16 good, 11 fair, 2 poor) as displayed by Fehler! Verweisquelle konnte nicht gefunden werden..

Figure 3: Overall performance referring to the criterion appropriateness



However, with regard to the validity of drawn conclusions, the quality of the evaluation reports deteriorates. In a quarter of the reports (7) inappropriate conclusions are derived from a methodological point of view. 3 reports are even rated very poor with regard to this issue. Even though 11 perform fairly and 11 well against WVG's requirements, CEval has to raise attention to the fact that most evaluation designs requested and agreed upon with WVG can be only second-best solutions given time and budget constraints and do not comply with scientific standards of validity (which will be further discussed in 4.4 and 4.6).

#### 4.4 Methodology

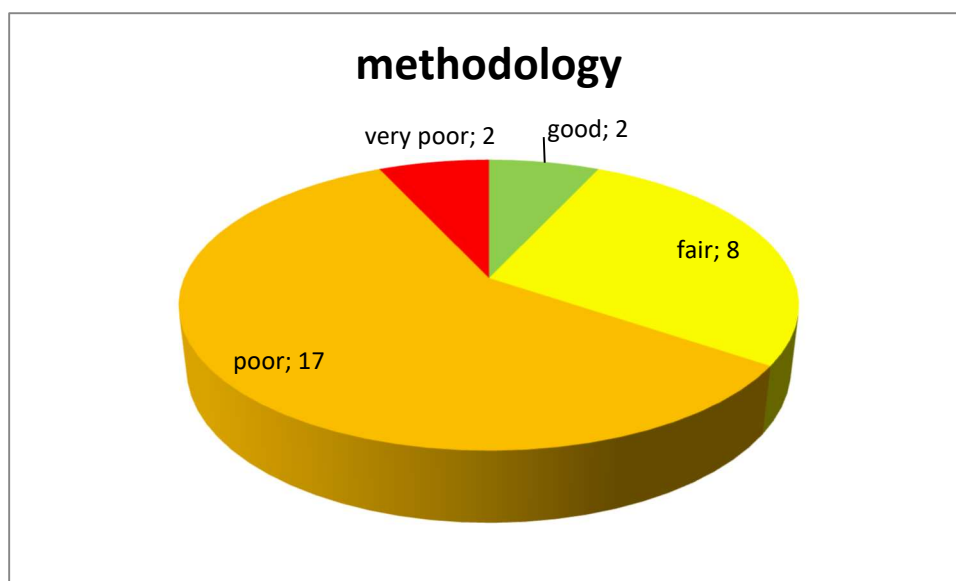
The methodology criterion is quite far-reaching as it does not only focus on the articulation of the results chain, but also emphasises on data sources, data collection, and analysis methods. Thus, it is no surprise that given the scope of this criterion, the performance of the evaluation reports vary widely according to these sub-criteria. While 85% of the reports do not adequately outline a results chain and programme theory, 50% specify sampling and analysis methodologies in at least a satisfying manner.

Also half of the reports fail not only to critically present the rationale for their quantitative and qualitative sample selection but also to discuss limits of the applied methodology.

Beyond these fundamental methodological concerns, it has been analysed to which extent the evaluation reports make reference to the capability and the robustness of the evaluated ADP's monitoring system. Thereby, the meta-evaluation discloses that the majority of reports (20) fails to address this aspect appropriately.

At a glance **Fehler! Verweisquelle konnte nicht gefunden werden.** reveals the assessment of the methodological performance of the evaluations. Whereas 10 reports comply to a sufficient extent with WVG's methodology criterion (2 good, 8 fairly), 19 reports do not comply with this quality standard (17 poor, 2 very poor).

Figure 4: Overall performance referring to the criterion methodology



A detailed view at the reports discloses further, that evaluation teams often fail to particularly mention and explain the usage of WV’s quantitative and qualitative data collection tools or even lack awareness regarding the wide range of WV’s instruments and therefore did not apply them at all. However, compared to the 2014 Meta Evaluation the use of the quantitative instruments is significantly higher. Especially the Functional Assessment of Literacy Tool (FALT) is frequently used to measure the literacy level of school children. The same holds true for the caregiver survey and the youth health behaviour survey (YHBS). The caregiver survey tool is applied in nearly half the evaluations carried out within the last two years. The application of the YHBS also increased substantially from 15 to 38%.

Table 1: 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	
	N	In%	N	In%
Functional Assessment of Literacy Tool (FALT)	5	15%	21	72%
Development Asset Profile (DAP)	3	8%	7	24%
Caregiver Survey	6	18%	14	48%
Youth Health Behaviour Survey (YHBS)	5	15%	11	38%
Measuring child growth	8	24%	9	31%
<b>Application of innovative qualitative methods</b>				
Comparison discussion group	4	12%	0	0%

Photo-voice	2	6%	1	3%
Seed assessment	5	15%	2	7%
Ladder of life	3	8%	4	14%
Tree of change	7	21%	7	24%

Apparently standard qualitative instruments like focus group discussions or key informant interviews are used on a regular basis. However, as Table 1 shows only a minority of reports used innovative qualitative methods like for example the tree of change or the ladder of life. In particular, the evaluation report of the Mannargudi ADP in India combined common qualitative data collection tools with non-traditional ones and moreover presented a comprehensive application of the tools.

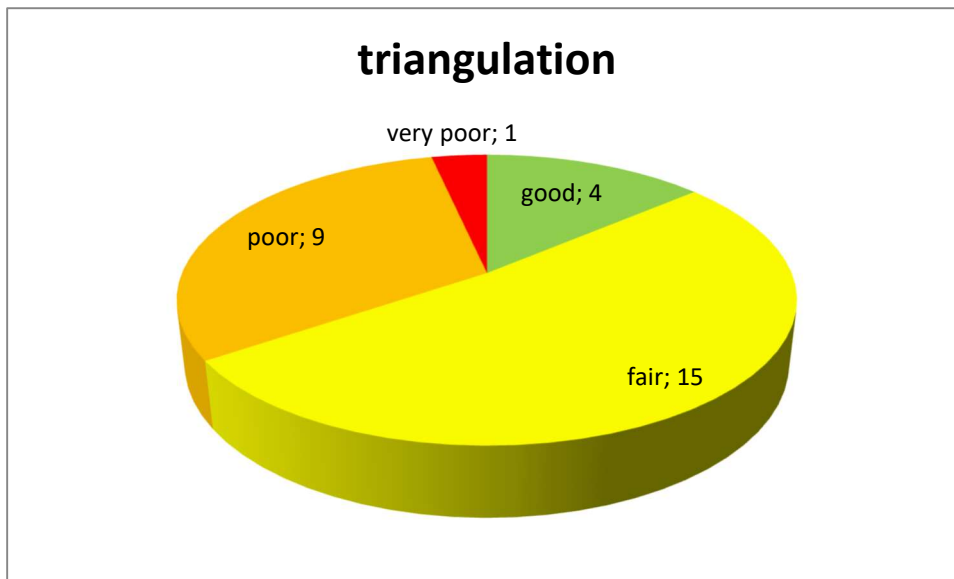
Unlike in the document analysis, where appropriateness of methods, methodology, and triangulation have been assessed separately, in the online survey there was only the question on the quality of methodology included in order to keep the questionnaire manageable for the respondents. While two third of the respondents, rate the quality at least good, also 17.5% see a clear need for improvement. Again, the LEAP guidelines are mentioned by several participants to be very precise and helpful on this issue. Nevertheless, several comments contain criticism concerning the rigorousness of data analysis. Apart from the potential for improvement in analytical skills, qualitative data furthermore revealed that the concept of sampling and representativeness is not clear among all WV evaluation team members. It is often opted to only collect data among the beneficiaries since including non-beneficiaries is considered to be way to time and resource intensive. These findings lead to the conclusion that in some cases the necessary methodological knowledge is lacking and hence, a sound data collection and analysis process is hindered. Although, the criterion is rated much better than in the document analysis the qualitative data clearly confirms that there is need for improvement as detected in the document analysis.

#### 4.5 Triangulation

This section shows to which extent the ADP evaluations triangulate data, methods and perspectives in their analysis. According to **Fehler! Verweisquelle konnte nicht gefunden werden.**, the overall picture for WV's triangulation is quite satisfying. While more than half of the reports comply with WV's standards mostly fair (4 rated good, 15 fair), only 10 fail to do so (9 rated poor, 1 very poor).



Figure 5: Overall performance referring to the criterion triangulation



A detailed look at the application of different data collection methodologies and data triangulation, however, provides a more positive picture as 23 out of the 29 reports use a method mix and a broad data base (4 rated poor, 2 rated very poor). Yet, this performance deteriorates when it comes to appreciating different perspectives of various stakeholders to explain how changes have occurred. 9 reports satisfy this sub-criterion in a good and 9 in a fair manner, while 10 are rated poor and one very poor.

Although the vast majority is on a state-of-the-art level regarding data collection, quality decreases when it comes to the appropriate discussion of results. Thus, 20 of the evaluation reports turn out rather poor in disclosing and interpreting apparently contradictory results (9 rated fair). Once again, given the broad heterogeneity among the reports, also positive examples can be found, such as the evaluation report of the Achuapa ADP in Nicaragua, which performs well with regard to WV's triangulation criteria. It provides a mixed-methods approach for triangulating qualitative data, including the tree of change method, development asset profile, key informant interviews, focus group discussions and site visits with quantitative data emanating from a household survey as well as with insights from reports published by official government institutions, and hence, includes several stakeholder perspectives.

#### 4.6 Identification of WV's contribution

This section illustrates to which extent the evaluation reports identify WV's contribution to observed changes in the working fields of the ADPs. While nearly half of the reports tackle this issue (6 rated good, 7 fair), the other half fails to explore a clear link between ADP's actions, its outputs and following outcomes (11 rated poor, 5 very poor).

To ascertain changes quantitatively it is important to use reference points. The comparison with baseline data qualifies for observing differences after ADP's interventions. By additionally using data of a comparison group – a group which is similar to the benefited one but which has not been affected by ADP's interventions – eventually allows isolating the impacts of an ADP from external factors

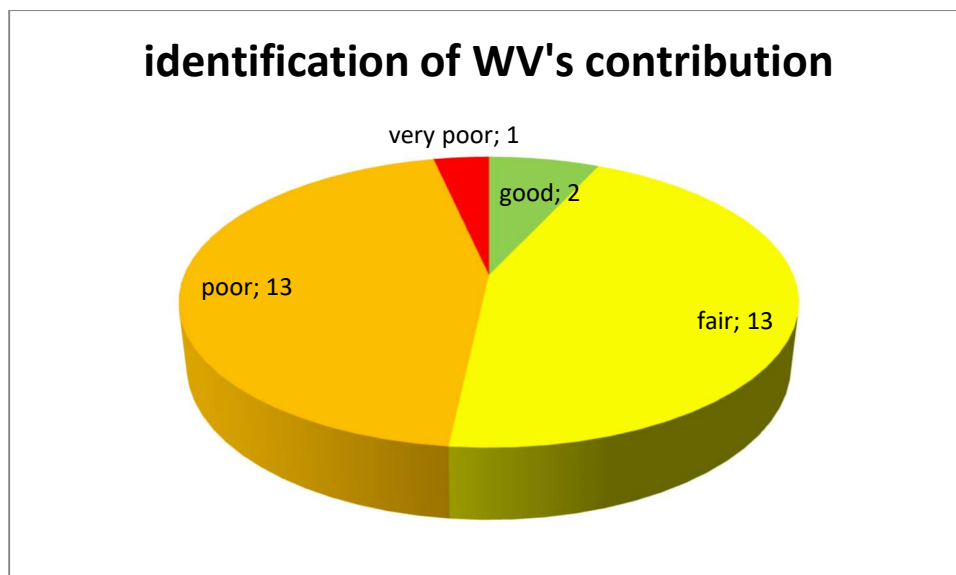
(i.e. infrastructural improvements, natural disasters, governmental development programmes), which also yield to change in the project area.

While only 6 of the 29 evaluations do not use a point of comparison at all, 7 evaluations include either baseline or comparison group data. While the remaining 16 reports also present some point of comparison, yet, the reader is either left with doubts if the comparison group is valid because its selection is not properly described or the sample size is rather small. Furthermore, baseline data is often not restricted to the intervention area (but to country level), contains non comparable indicators or is not representative. Hence, the contribution of the ADP has not been explored rigorously.

The reason why overall this criterion is still much better evaluated than in 2014 is because alternative factors, which could also have contributed to observed changes are considered far more often when interpreting collected data. Furthermore, positive as well as negative unintended and unexpected changes are not completely absent in the evaluation reports anymore (in 2014 none of the evaluation reports considered unexpected changes). However, there is a clear need for improvement in both areas. Only 9 reports make reasonable (qualitative) considerations on alternative factors (4 are rated good, 5 fair), which means that nearly 70% do not consider external factors in a satisfying manner (12 are rated poor, 8 very poor). A good example in that regard is the Mukumbura ADP in Zimbabwe which highlights influences from governmental and other non-governmental organizations detected during focus group discussions. Finally, references to unintended or unexpected changes (negative or positive) are available in 12 reports while 60% (17 reports) still overlook these.

**Fehler! Verweisquelle konnte nicht gefunden werden.** shows the overall compliance with WV's contribution criterion which reveals that still nearly half of the reports did not manage to attribute observed changes to ADP's interventions.

Figure 6: Overall performance referring to the criterion contribution



Results from the online survey confirm the still existing need for improvement. Of all requested criteria this is the one where most of the staff members see a need for improvement. While all other criteria have only been evaluated with substantial need for improvement by below 5% of the participants (except for methodology with 6.3%), 11% do see a need for improvement, 14.3% even a substantial need for improvement in this area. For around 20% this criterion is assessed satisfactory. As

detected in the document analysis, the qualitative data from the online survey shows that this is mostly due to the minor quality of baseline data or even missing baseline data. All participants who commented their rating also mention the lack of a control group as a reason for insufficient identification of the own contribution. Still more than 50% rate the quality of this criterion as at least good. Nevertheless, when prioritizing the most important needs for a high quality evaluation 42.6% name the evaluation design as the most important aspect. Only „qualification of the evaluation team“ is chosen slightly more often (45.1%).

#### 4.7 Satisfaction of information needs

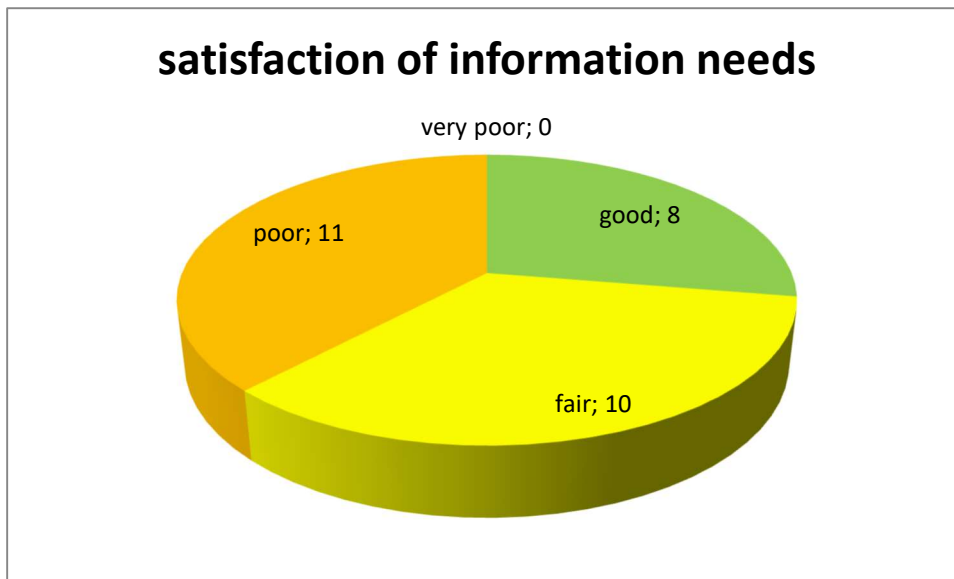
The same methodology as used for WVG's criteria is applied for the two additional criteria, which are assumed to be valuable for WVG to further precise the Terms of References (ToR) of future evaluations and thus, to improve the benefits and the quality of evaluations. CEval examined the criterion "satisfaction of information needs" to inquire to which extent a particular evaluation provides WVG with appropriate recommendations and lessons learnt. Moreover, the criterion "organisation of findings" is proposed to highlight whether the reports follow a clear structure and are embedded into a broader theoretical concept.

The analysis reveals that many evaluation teams fulfilled the ToR in an acceptable manner, while only a few did not complete the assigned tasks accordingly (11 reports rated good, 12 fair, 5 poor, 0 very poor). A look at the recommendations shows that the vast majority of the reports (13 rated fair, 8 good) succeeds in providing appropriate suggestions on how to continue with an ADP or in case of a final evaluation; what to consider in future ADPs. 18 reports went beyond deriving recommendations and provide lessons learnt during the evaluation to share with WVG and thus, to enhance chances of improvement in future.

Due to its importance the online survey specifically asked for the quality of recommendations. Two third consider the given recommendations as good or very good (44.4% good, 22.2% very good) while 11.1% assess it as fair and only 11.1% see a need for improvement. The percentage of staff, which sees a substantial need for improvement, is below 2%.

Overall assessment as displayed in **Fehler! Verweisquelle konnte nicht gefunden werden.** is thus predominantly promising. Although 11 reports cannot satisfy WVG's information needs (rated poor), none is rated very poor. 10 reports provide a fair amount of information needed and the remaining 8 a good amount.

Figure 7: Overall performance referring to the criterion satisfaction of information needs



If the information in the evaluation report covers the needs in a satisfactory manner can best be answered by the team who developed the ToRs as well as the persons responsible for program design. Hence, WV staff itself. Nevertheless, staff members rated this criterion based on their different experiences and expectations and furthermore did not specifically rate the reports presented for the document analysis. This might be the reason why in the document analysis only less than a third of the reports satisfy the information needs in a good manner, while half of WV's staff claims that it assesses this criterion well, 16% even very well.

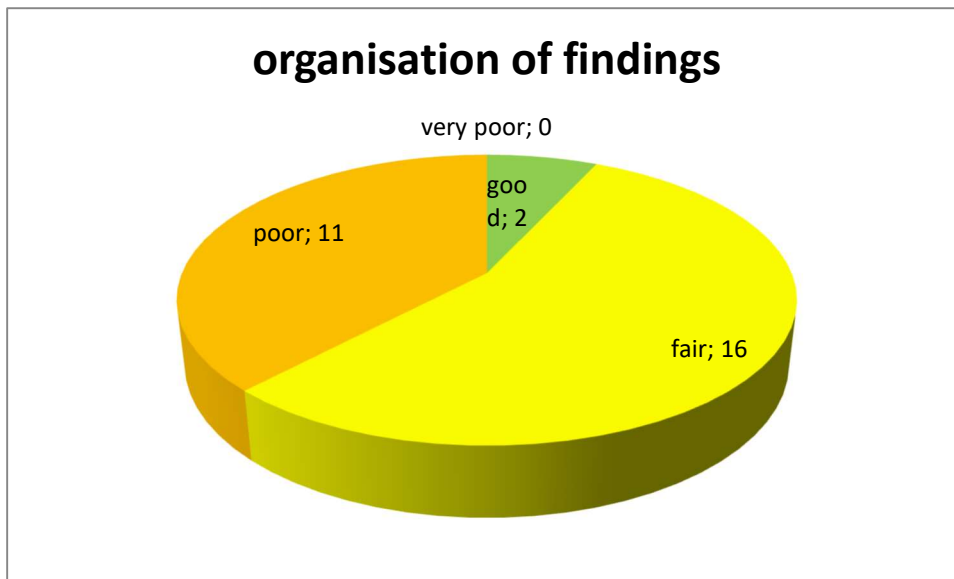
#### 4.8 Organization of findings

It is one challenge to conduct an evaluation, which complies highly with WVG's evaluation criteria. Presenting the results of an evaluation in an excellent manner and conceptualise findings accordingly are additional challenges. 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.

While the ToRs often specify (partly very detailed) how evaluation reports have to be structured, they are not very explicit on how to summarise findings. Although it is positive that many reports contain an executive summary, it would be helpful to find tables summarising main results which should be organised according to the ADP's log frame indicators. While nearly a third of the reports fail to do so, the ones who present summarizing tables do it in a satisfactory manner (3 poor, 6 fair, 11 good). The same holds true for general structure. Most of the reports are organized along the logframe indicators (11 are rated good, 8 fair, 6 poor and 4 very poor according to this sub-criterion) and/or according to DAC criteria.

A closer view discloses that beyond organising findings, most evaluation reports completely fail to embed the analysis into a conceptual framework. 24 of the reports lack explicit statements on how the interventions of an ADP impact on the beneficiaries or put differently through which transmission channels change was produced (2 rated good, 2 fair, 19 poor, 5 very poor). Nevertheless, the overall picture is quite satisfying since all sub-criteria are weighted equally for the aggregated rating.

Figure 8: Performance referring to organisation of findings



Organization of findings is rated mostly “good” by WV staff (58% good, 8% very good). However, about 19% think that the organization of findings is fair, while 13% see a need for improvement and a really small percentage (2%) even sees substantial need for improvement. Hence, findings are similar to the ones from the document analysis. Qualitative data shows, that the LEAP templates are a very useful guideline. Nevertheless, as also noticed during the document analysis, WV staff often mentions the huge differences of the presented evaluation reports. Although most of them follow the LEAP guideline, the understandings, capacities as well as aspirations, evaluands and evaluation teams are very different.

After presenting the results of the document analysis and linking them to the perception of the WV staff, the further findings of the online survey will be presented.

## 5. Further findings from the online survey

Apart from the questions which have been included into the online survey to check for congruent tendencies of the findings from the document analysis with the perception of WV staff, additional questions concerning the quality of evaluations and questions concerning the usage, utility and usefulness of evaluation reports within WV have been included.

Participants were asked to choose and rank the three most important factors influencing the quality of an evaluation. Possible choices were: available time, budget, qualification of the evaluation team, evaluation design and evaluation instruments and methods. As already mentioned before, the most important factors influencing the quality of an evaluation according to the participants of the online survey, are the qualification of the evaluation team, and the evaluation design. Also the time available for an evaluation is considered a key factor by a third of the participants whereas evaluation instruments and the budget mainly occupy secondary or third rank.

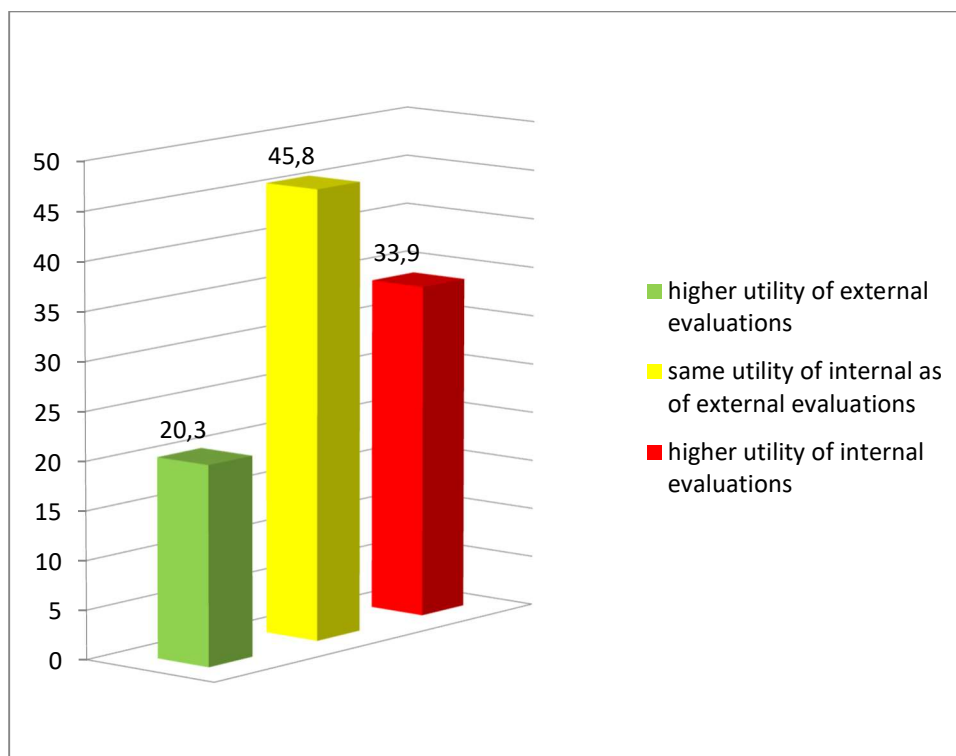
Concerning the general quality of the evaluations 57.4% of the participants perceived a slight improvement within the last two years, 26.2% even a substantial improvement. The percentage of par-

ticipants who noticed a decline in quality is below 5%. The positive development is congruent with the results from the document analysis.

Concerning the utility of evaluations, qualitative data analysis shows that all participants agree that evaluations serve the improvement of project re-design and future project design. Although one would expect that mid-term evaluations are assessed as more useful than final evaluations because they allow to directly apply the new insights onto the same context with the same stakeholders, this is not the case. The percentage in favor of mid-term evaluations is equal to the percentage of final evaluations (17.7%). More than half of the sample (56.5%) does not have any preferences but considers both types of evaluations as equally useful.

The biggest part of WV's evaluations is conducted internally. The results of the online survey show that there is no clear preference for either internal or external evaluations. Nearly half of the participants consider both as equally utile, while 20.3% consider external and 33.9% internal evaluations as more utile (see Figure 9). Although, internal evaluations have certain advantages (i.e. less cost intensive, working up to own expectations, deep understanding of intervention etc.) they also bear biases and hence, compromise on objectivity. Thus, a compromise could be an internally lead but externally implemented evaluation, since it is especially important that interviews are not conducted by WV members themselves.

Figure 9: Utility of internal and external evaluations



Asking for examples where evaluation results have been used, most participants refer to the incorporation of the results in the Child Well Being Reports<sup>9</sup>. Furthermore, the evaluation findings have to be presented to the support offices and to the donors. Additionally the qualitative data analysis shows that there is great consciousness for the importance of the distribution of the results. Participants

<sup>9</sup> Child Well Being Reports are published by WV since 2011. They offer an overview about the achieved goals using specific child well being outcomes defined by WV ([http://www.wvi.org/sites/default/files/Child\\_Well-being\\_Aspirations\\_and\\_Outcomes\\_English\\_0.pdf](http://www.wvi.org/sites/default/files/Child_Well-being_Aspirations_and_Outcomes_English_0.pdf))

often mention, sharing the results with governmental institutions as well as other non-governmental organizations to raise incentives for similar projects or to positively influence already existing or upcoming projects of these organizations. Furthermore, like previously mentioned, WV staff strongly opts for discussing the evaluation results with the target group to encourage beneficiaries to continue with the different implemented projects and hence, assure sustainability.

The additional results from the online survey show that no matter if intern or extern, evaluations are useful and actually used. Results are considered in the learning process. Thus, they are mostly used to steer and influence future projects. Furthermore, experiences are often considered in the project planning of other institutions.

## 6. Conclusion

The overall assessment of the evaluation reports against WVG's evaluation quality criteria leads to a quite heterogeneous picture as shown in the table below.

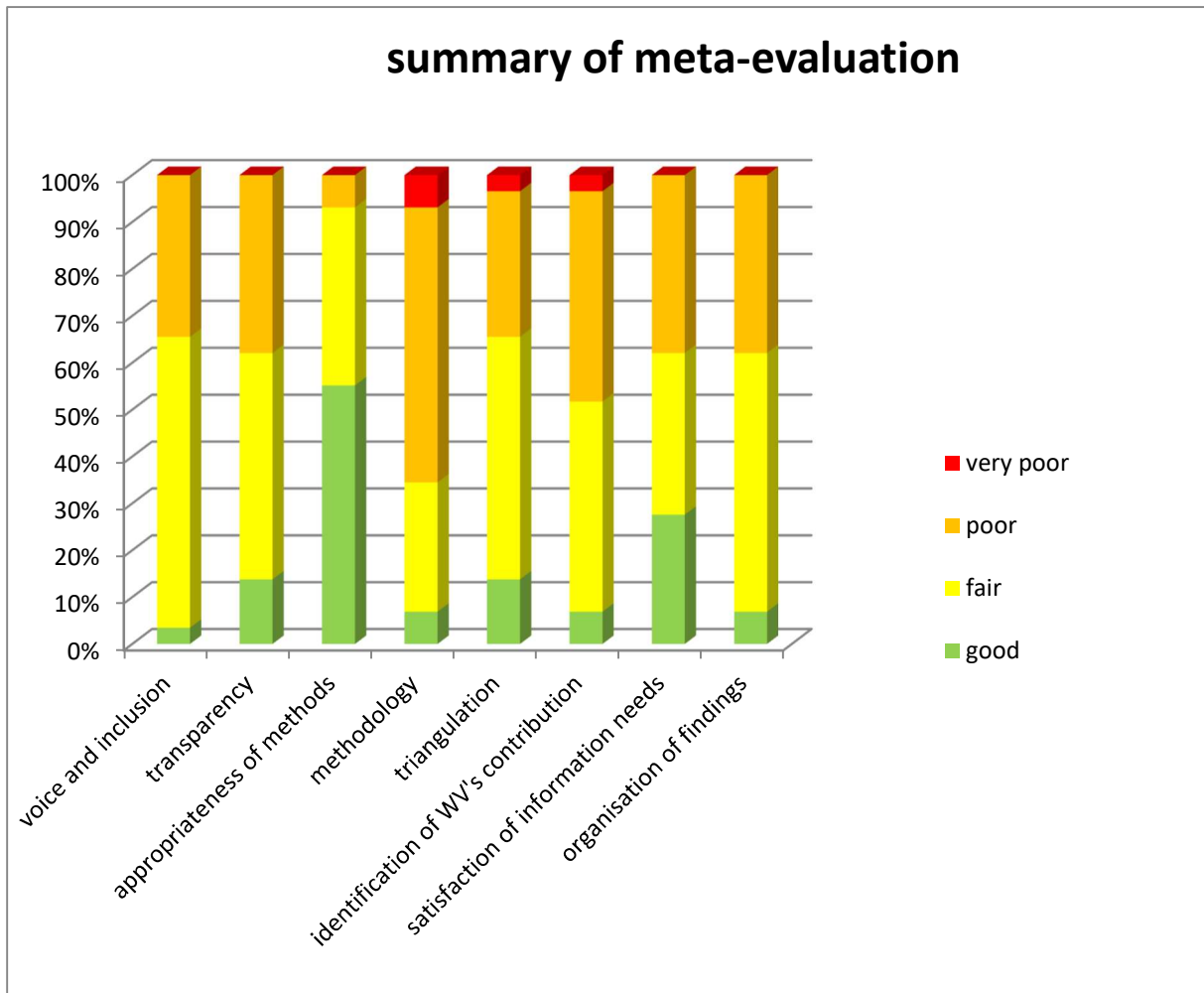
Table 2: Evaluation results matrix

Evaluation criteria by ADP	Armenia (Vardenis) Cambodia (Rovieng Area) Cambodia (Tbeng Meachey) Georgia (Abkhazia) India (Mannargudi) Kenya (Mutonguni) Kenya (SAPLIP) Lebanon (Kds) Lebanon (Saaf Consult) Malawi (Kamwendo) Malawi (Mkumba) Mali Mongolia (Zunkaharaa) Mozambique (Saua Saua) Nicaragua (Achuapa) Nicaragua (Sauce) Philippines (North-West Leyte) Senegal South Africa (Ekukhanyeni) Swaziland (Gilgal) Swaziland (Matsanjeni) Tanzania (Magoma) Tanzania (Mzaki) Vietnam (District 8) Vietnam (Trier-Phong) Vietnam (Vinh Linh) Zimbabwe (Chipi) Zimbabwe (Mukumbura) Zimbabwe (Rushinga)
voice and inclusion	
transparency	
appropriateness	
methodology	
triangulation	
contribution of WV's intervention	
satisfaction of information needs	
organization of findings/conceptualization of change	



The results matrix shows that many reports reveal shortcomings. Ordering the reports chronologically and clustering them regionally (Africa, Asia, Latin America, Eastern Europe) did not disclose any systematic differences. Hence, figure 10 highlights average strengths and weaknesses of the ADP evaluation reports best.

Figure 10: Summary of meta-evaluation results



The results of this year’s meta-evaluation feature some similarities to those of the meta-evaluation that was conducted two years ago. About half of the evaluation teams select highly appropriate methods and apply them consistently, and slightly more than half of the reports satisfy the information needs well or fairly. Furthermore, voice and inclusion, transparency, triangulation and organization of findings are mostly rated fairly and hence, still present potential for improvement. Finally, with regard to the quality of the methodology the evaluations perform rather poorly. Thereby, the lack of a results chain is one of the main reasons why the methodology criterion is rated poor. With regard to the identification of WV’s contribution some improvement in comparison to the last meta-evaluation is visible as baseline data appears to be included increasingly. Yet due to lacking comparison group data it is still mostly not possible to provide for a reliable impact attribution<sup>10</sup>.

<sup>10</sup> 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.

Nevertheless, a positive trend can be clearly noticed comparing the overall results of the two meta-evaluations. The results from the online survey show that also WV staff has noticed an improvement in quality over the last two years.

The online survey revealed that evaluations in general, regardless of being mid-term or final evaluations, internal or external evaluations are considered as useful by the majority of the participants. Results are included into the learning process and used to improve future projects. Moreover, experiences are shared with other institutions and thus also influence their project planning.

This meta-evaluation disclosed that WVG's ADP evaluations are not yet state-of-the-art. While, a positive trend could be detected, WV should consider the following recommendations to further improve its overall evaluation performance.

## 7. Recommendations

To optimize the evaluations with regard to the voice and inclusion criterion, the reports should outline more information on how and up to which level beneficiaries and partners have been included during the intervention. Furthermore, beneficiaries would have to be considered more during evaluation design and analysis. However, WVG may rethink to which extent participation of beneficiaries and other stakeholders will be de facto feasible and meaningful at the analysis stage since methodological standards have to be met and objectivity is not guaranteed. Nevertheless, results should be disaggregated according to relevant social differences and definitely be presented to the target group to not only foster the acceptance of the given recommendations but also motivate the beneficiaries to continue with the projects and hence, have a positive effect on sustainability.

To ensure objectivity it is recommended to mandate independent experts with the implementation of the evaluation or at least to involve them as methodological advisors. In order to prevent biased results it is not only important that the lead consultant is not on the payroll of WV, but also the enumerators or facilitators of focus group discussions. Furthermore, it has to be disclosed in the report if the evaluation has been conducted internally or externally. The LEAP Design does not intend to put the logos of external consultants on the reports. However, this would help to directly identify if the evaluation has been externally implemented, accompanied or conducted fully by WV staff.

To ensure that the causal assumptions according to which the respective project is evaluated are clear to all readers, it is advisable that the results chains and/or programme theories are outlined in the report.

Another aspect, that leaves room for improvement is the validity of the evaluation results. Validity refers to the extent to which an analysis really measures what it attempts to measure. As detected in Chapter 4.6, the majority of the evaluation reports do not yield to valid results as they fail to detach the net effects of an ADP's intervention from observed gross outcomes in the programme area. To enhance validity WV should revise their evaluation strategy and consider pooling resources to conduct less but therefore methodologically more sophisticated evaluations, e.g. based on quasi-experimental designs, that allow for impact attribution and identifying the project's net effects.

In any case where quantitative data is collected, the sampling approach should be explained in order to allow the reader to assess the representativeness of the survey results.

However, where sophisticated evaluation designs are not feasible, it is even more important to point to limitations of the collected data, methodology and results and particularly to discuss descriptive statistics in a careful way, e.g. by not only presenting mean values, but also providing further information on the explanatory power of the data. Moreover, qualitative data should be used for providing plausibility about WV's contribution to observed changes. For that the usage of WV's innovative qualitative data collection instruments could be highly useful since some of them are developed to obtain information from specific target groups, i.e. children or juveniles. Furthermore, qualitative data often gives hints about possible external factors. These should be considered when discussing the results.

Although the LEAP turns out to be a good guideline it is quite extensive and theoretical. Practical examples could help to improve the overall quality of evaluation reports. Therefore, CEval suggests increasing the exchange of evaluations. Establishing a database containing promising examples would be a good source to get access to practical information for further development or future design of projects.

## 8. Annex

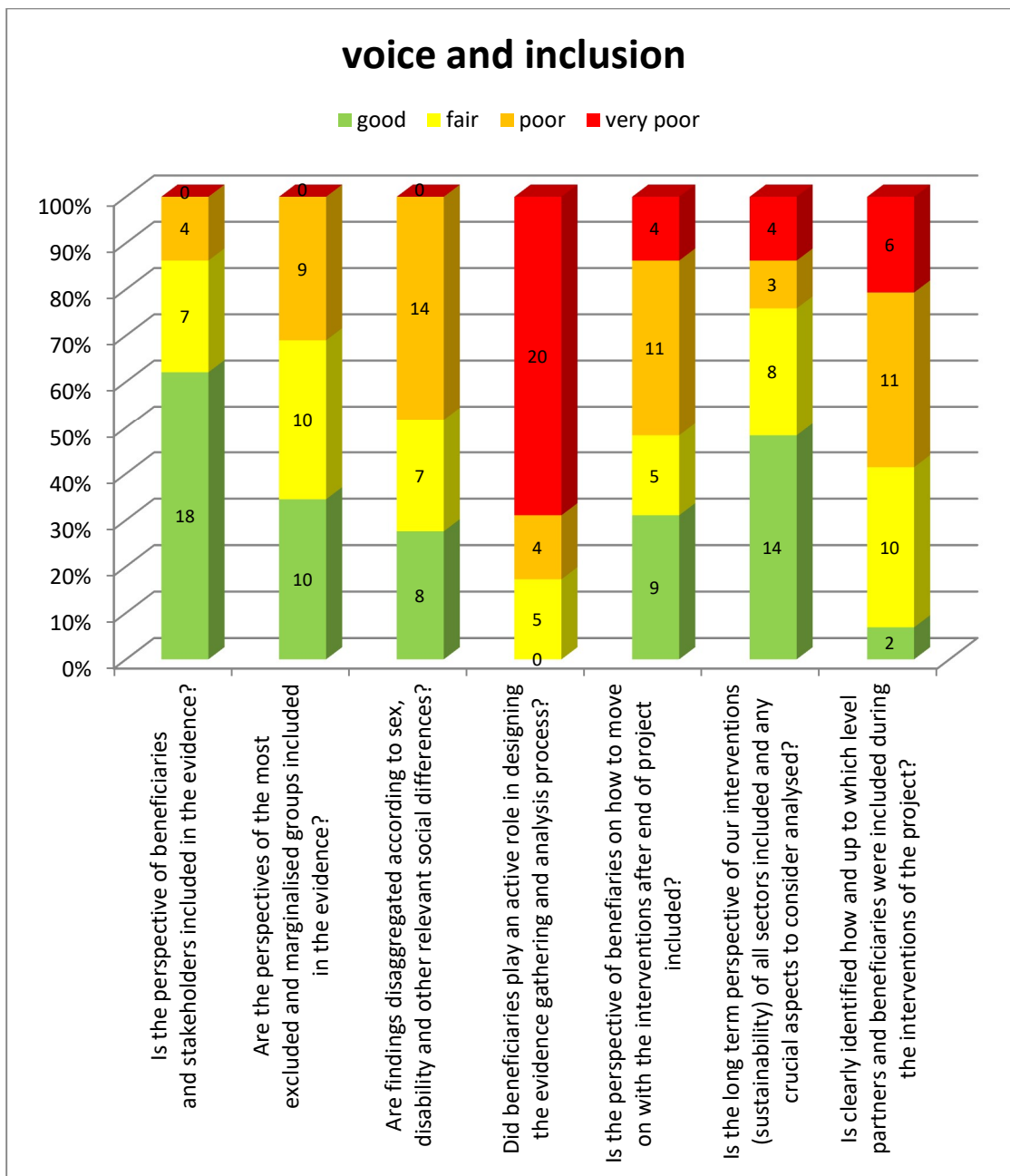
### 8.1 Criteria and Sub-criteria



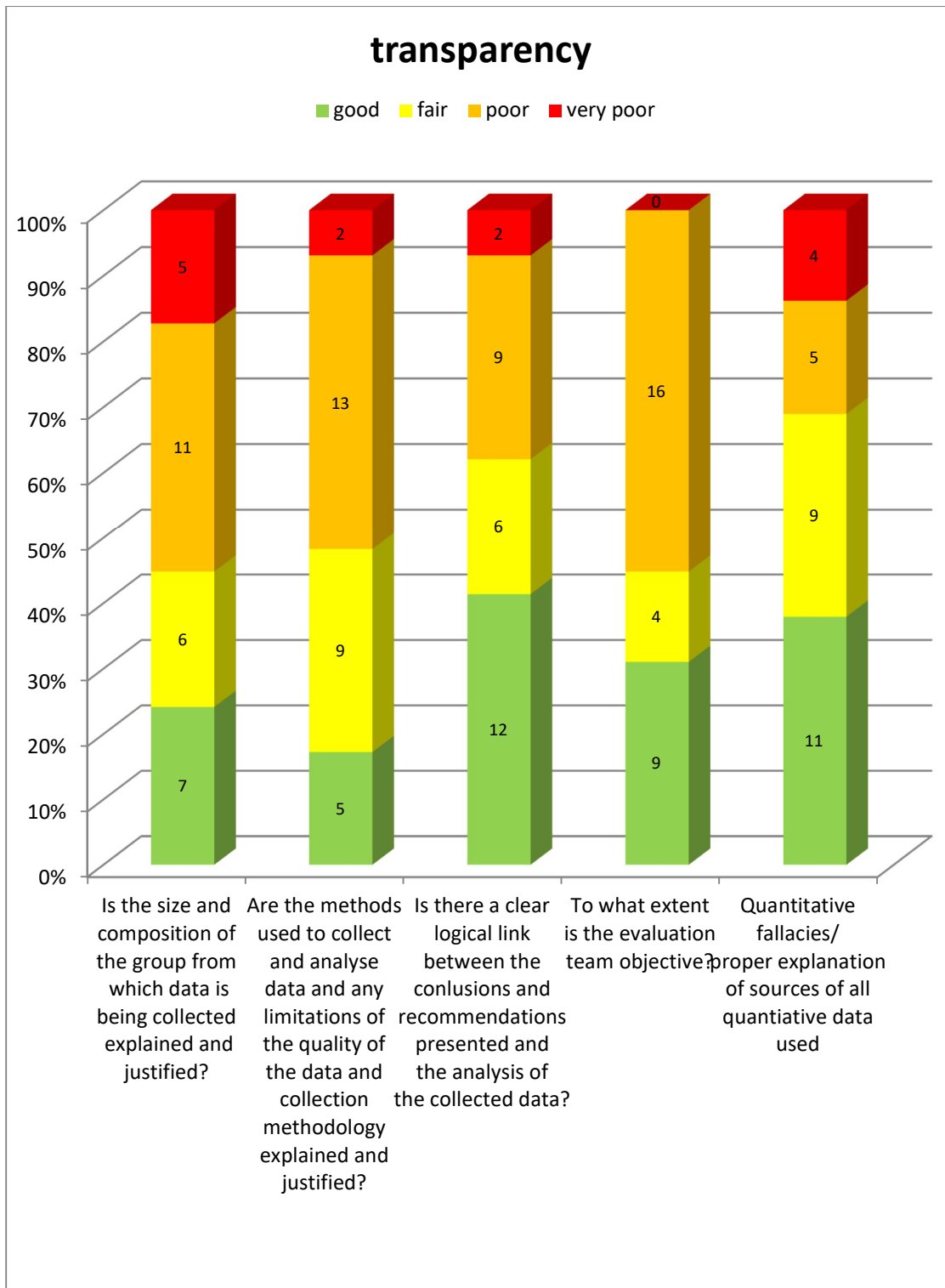
Subcriteria.xlsx

### 8.2 Graphical Illustration of Sub-criteria

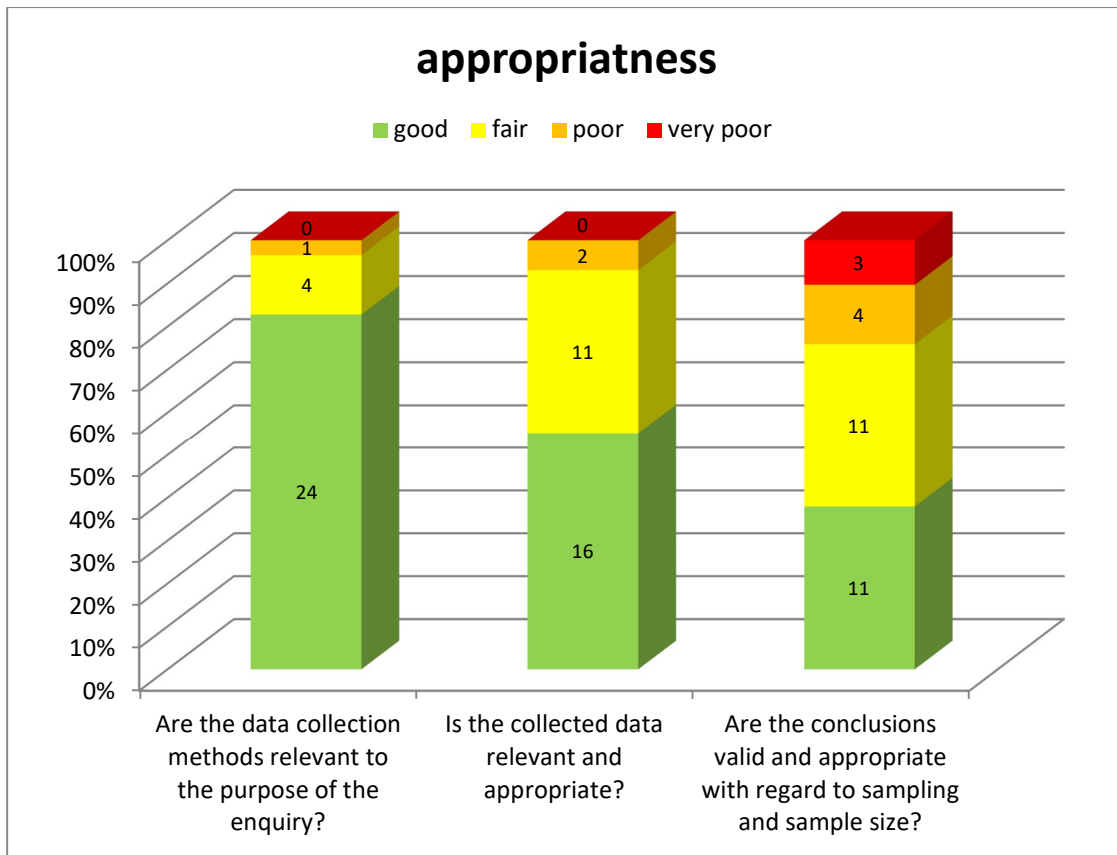
#### 8.2.1 Voice and Inclusion



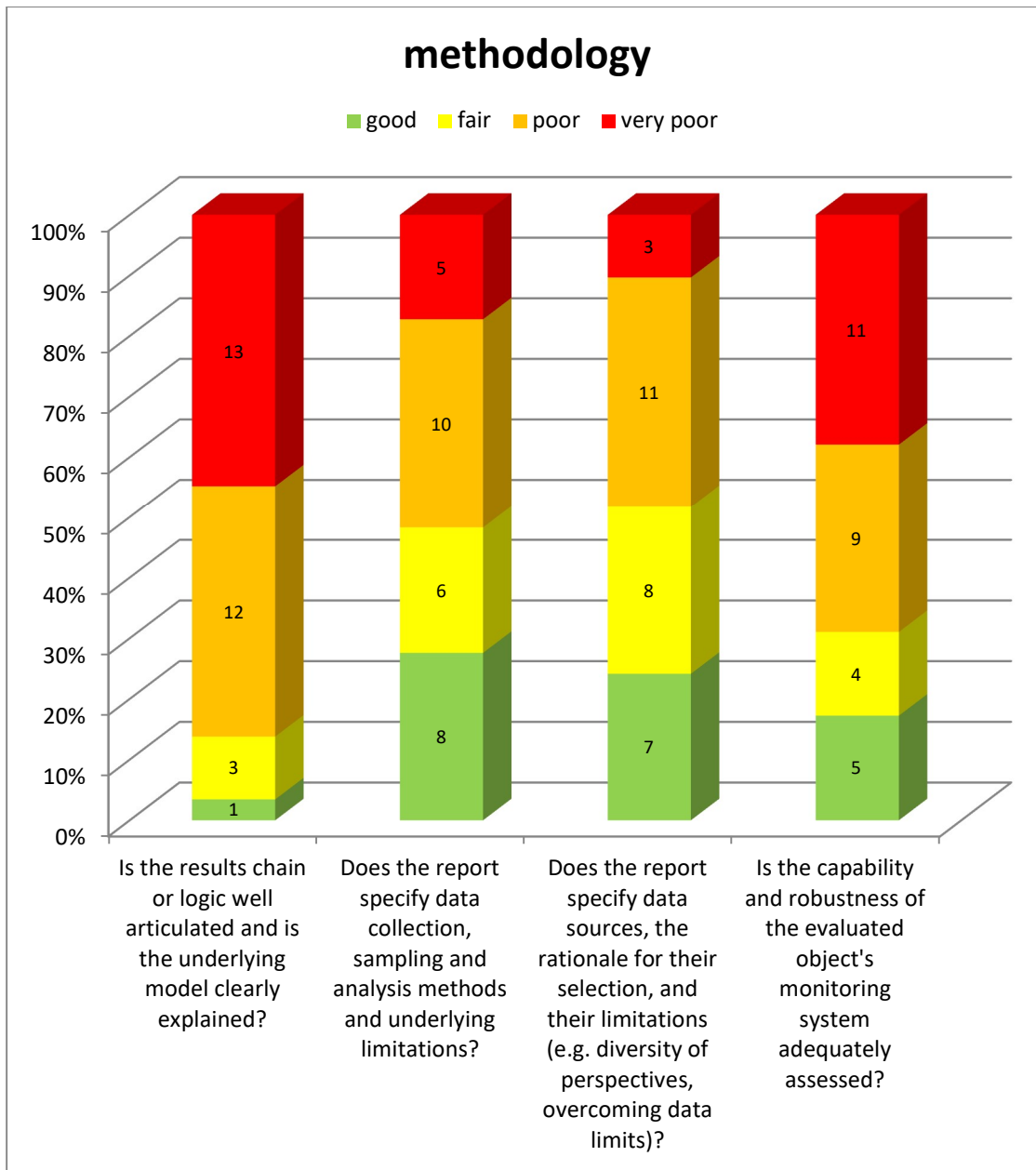
8.2.2 Transparency



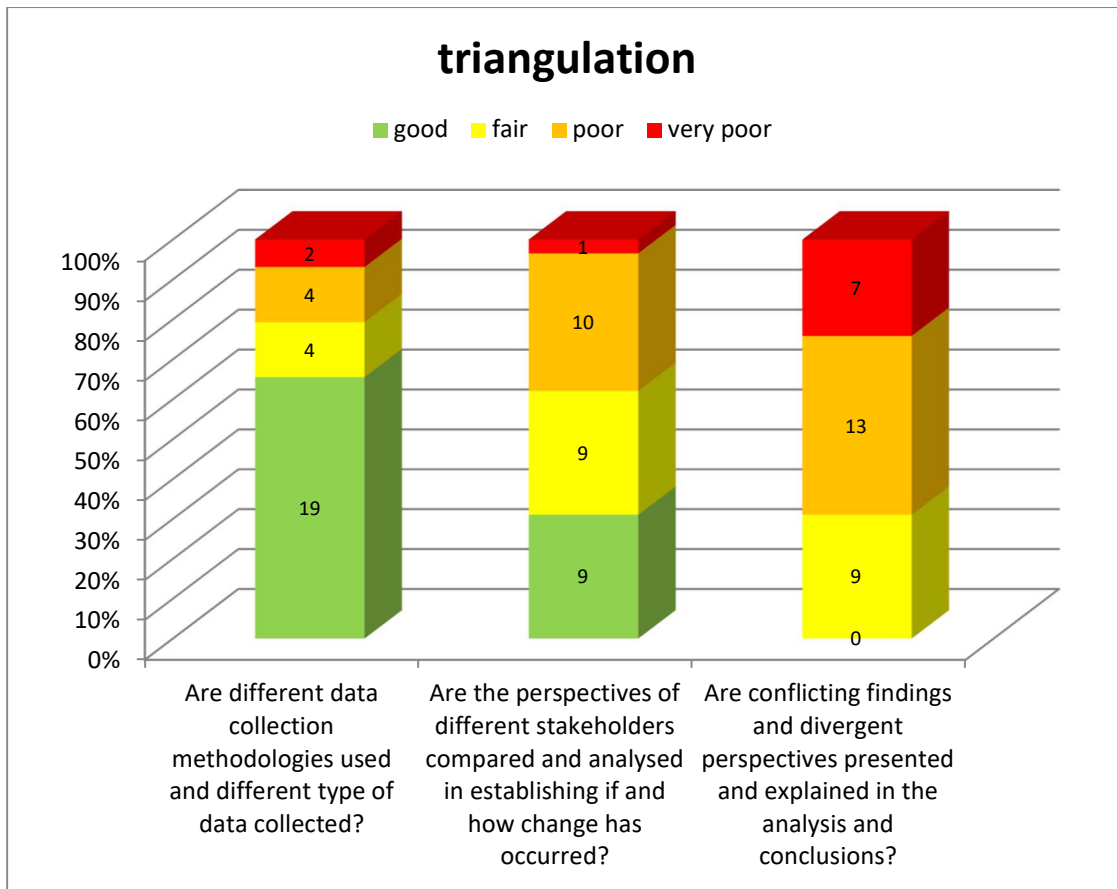
8.2.3 Appropriateness



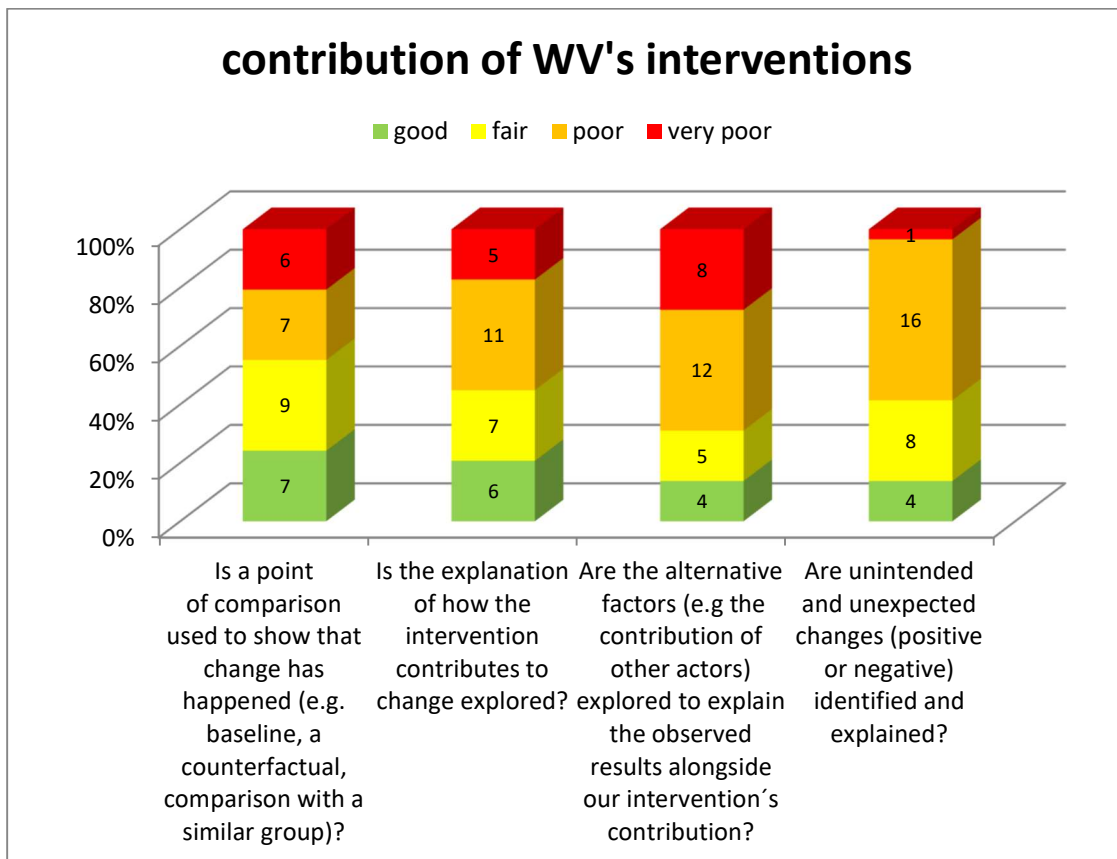
8.2.4 Methodology



8.2.5 Triangulation

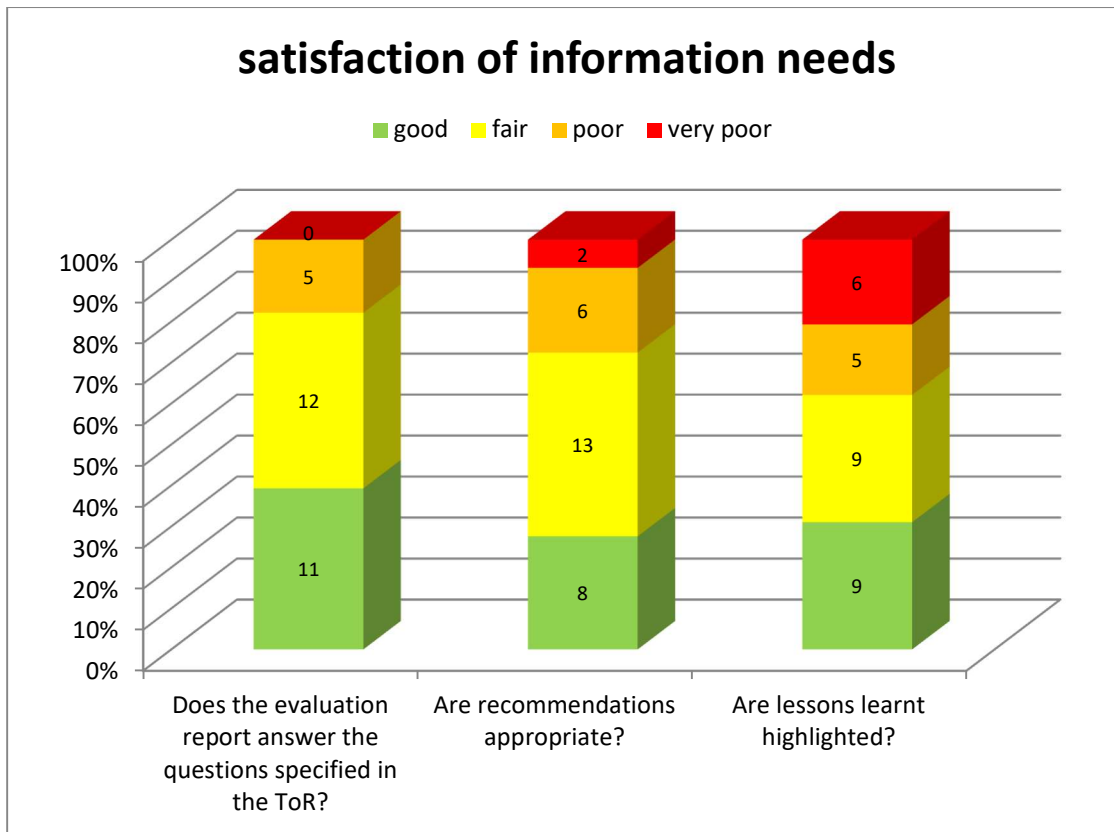


8.2.6 Contribution of WV's interventions





### 8.2.7 Satisfaction of information needs



\*ToR for one report was not available, hence the first sub-criterion could not be rated.

### 8.3 Online-Survey Questionnaire



Questionnaire.pdf

### 8.4 Online-Survey Results



Results\_OnlineSurvey.docx