Meta-Evaluation of 34 evaluation reports of World Vision Germany

on behalf of World Vision Germany

Authors:
Susanne Johanna Väth
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.vaeth@ceval.de
URL http://www.ceval.de

Table of contents

1. Executive summary in German (Kurzzusammenfassung) 1
2. Background 2
3. Methodological approach 2
4. Assessment according to WVG's quality of evaluation criteria 3
4.1 Voice and inclusion of beneficiaries 3
4.2 Transparency 4
4.3 Appropriateness of evaluation methods 5
4.4 Methodology 6
4.5 Triangulation 8
4.6 Identification of WV's contribution 9
4.7 Additional criteria 10
5. Synthesis 11
6. Conclusion and recommendations 14

1. Executive summary in German (Kurzzusammenfassung)

Im Auftrag von World Vision Deutschland (WVD) hat das Centrum für Evaluation (CEval) eine Meta-Evaluation auf der Basis von 34 Evaluationsberichten zu WVDs langfristig angelegten Regional-Entwicklungsprogrammen durchgeführt. Um die Qualität der Evaluationsberichte zu bewerten, hat sich das CEval methodisch auf ein zweistufiges Auswertungsverfahren konzentriert. (Die Qualität der Programme selbst ist dabei nicht Gegenstand der Analyse). In einem ersten Schritt wurden die von WVD erarbeitete Kriterien Mitsprache und Inklusion, Transparenz, Angemessenheit der Methoden, Methodik, Triangulation, sowie Identifizierung des Programmbeitrags anhand von verschiedenen Unterkriterien bewertet, die schließlich in einem zweiten Schritt aggregiert wurden. Darüber hinaus hat das CEval die drei zusätzlichen Kriterien Befriedigung des Informationsbedarfs von WVD, Ergebnispräsentation sowie Konzeptualisierung der beobachteten Veränderungen eingeführt und ebenfalls in einem zweitstufigen Verfahren bewertet.

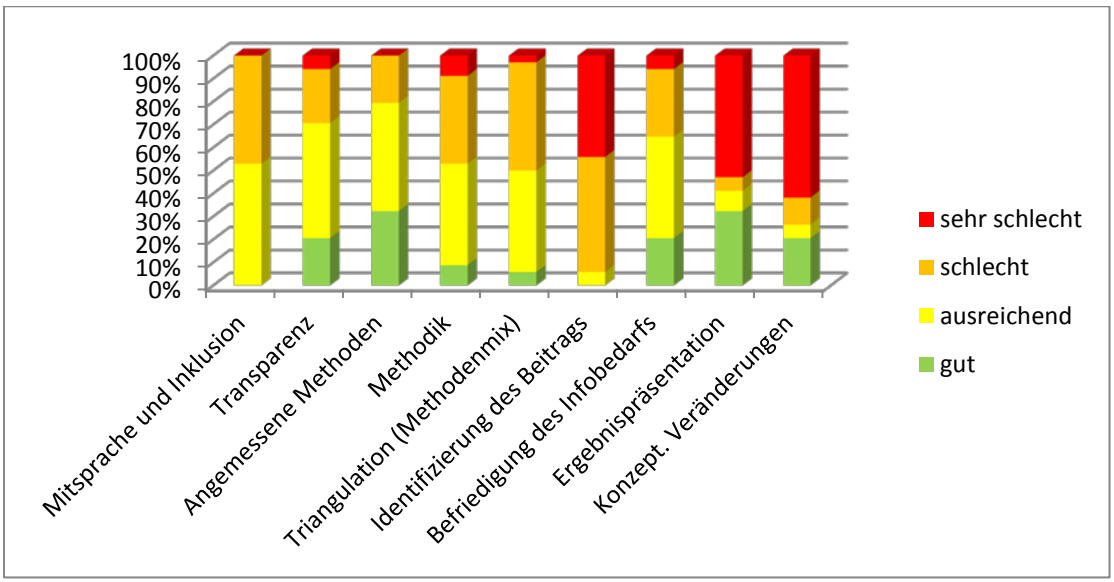


Abbildung 1: Zusammenfassung der Ergebnisse

Abbildung 1 zeigt, dass ein Großteil der Evaluationen noch nicht den gängigen Evaluationsstandards entspricht. Vor allen Dingen taten sich die Evaluationsteams schwer, den Programmbeitrag eindeutig zu identifizieren. Auch bezüglich der anderen Kriterien offenbart die Meta-Evaluation großes Verbesserungspotential. Ungeachtet des großen Verbesserungsbedarfs ist jedoch festzuhalten, dass in der Summe ein sehr heterogenes Bild entstanden ist. Im WVD vorliegenden Bericht der Meta-Evaluation kann das CEval dementsprechend auf einige Evaluationsberichte verweisen, die richtungsweisend für künftige Evaluationen sein könnten.

Um künftige Evaluationen zu verbessern schlägt das CEval unter anderem vor, dass WVD bereits bei der Erstellung der Ausschreibung festlegt, dass Evaluationsergebnisse entlang der Indikatoren eines Programms zu organisieren und entsprechend zusammenzufassen sind und dass die zu beobachteten Veränderungen explizit in eine adäquate Programmtheorie eingebettet werden müssen. Um den Programmbeitrag angemessen zu identifizieren bedarf es zudem einem vermehrten Einsatz von quasi-experimentellen Evaluationsdesigns, die über einen reinen Soll-Ist-Vergleich hinausgehen. Denn letztere erlauben nicht den Beitrag der Regional-Entwicklungsprogramme von den Beiträgen anderer Akteure (z.B. staatlichen Organisationen) oder anderer Faktoren (z.B. Umweltveränderungen) zu trennen und verbieten dementsprechend Aussagen über den Beitrag, den WVD zu den Veränderungen die sich über einen bestimmten Zeitraum im Programmgebiet beobachten lassen, geleistet hat.

2. Background

To maintain its programmes and activities a non-governmental development partner 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.

To assess the methodological soundness of its evaluation reports WVG engaged the Center for Evaluation (CEval) to conduct a meta-evaluation in the thematic areas of child well-being and community development. CEval thereby reviews the validity, the reliability, and the objectivity of the evaluation results against derived conclusions. As WVG developed (with impulses of CEval) a checklist for the quality of its evaluations, the assessment is structured according to its criteria:

1. Voice and inclusion of beneficiaries,
2. Transparency,
3. Appropriateness of evaluation methods,
4. Methodology,
5. Triangulation, and
6. Identification of WV's contribution.

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 (chapter 4), before a synthesis of the findings (chapter 5) follows. In concluding, this report offers recommendations to improve future evaluation designs and their implementation in accordance with the OECD/DAC Criteria for Evaluating Development Assistance.

3. Methodological approach

This meta-evaluation is based on 34 reports which have been produced between December 2011 and May 2014. They summarise the evaluation process of WVG's Area Development Programmes (ADP) in Sub-Sahara-African, Asian and Latin America countries, and are thus written in English (19), French (4) or Spanish (11). 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 four 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 15 years and were 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 available for all ADPs. 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 assess their quality against WVG's criteria.

In doing so, a checklist serves as 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 WVG's six criteria (i.e. voice and inclusion, transparency, appropriateness of evaluation methods, methodology, triangulation, and contribution). Furthermore, findings will be consolidated and one aggregated rating per criterion will be deviated.

The same methodology is applied for some 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 independent evaluations. First of all, CEval suggests the criterion "satisfaction of information needs" to inquire to which extent a particular evaluation provides WVG with appropriate recommendations and lessons learnt. Moreover, the criteria "organisation of findings according to log frame indicators" and "conceptualisation of change" are proposed to highlight whether the reports follow a clear structure and are embedded into a broader theoretical concept.

With respect to the limited scope of this meta-evaluation, CEval cannot assess each report in detail. This holds especially true for analysing the content in an all-embracing manner. Therefore, the focus lies on methodological issues, and on some 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. Further, for each report specific sections are assessed to inquire how well methods are applied and how valid findings are deviated. By doing so, this meta-evaluation identifies general trends, displays heterogeneity, and prepares the ground for enhancing the quality of ADP evaluation.

To ensure high quality of our work, one consultant leads the analysis and is in charge of reporting, while a second consultant ensures technical backstopping.

4. Assessment according to WVG's quality of evaluation criteria

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 (32 out of 34) perform well with regard to capturing the voice of a wide range of stakeholders, especially the report of the Ham Thuan ADP in Vietnam can be highlighted as a good example for future evaluations as it does not only use various data collection instruments to capture the voice of diverse stakeholders appropriately, but also highlights different perceptions of various beneficiaries. In contrast, only about half of the reports present disaggregated findings appropriately (7 rated good, 12 fair, 12 poor, 3 very poor).

Our analysis furthermore reveals that the beneficiaries seem not to play an active role in designing the evidence gathering and analysis process (32 rated poor, 2 very poor). Although, empirical data is

often gained in a participatory manner (i.e. by conducting focus group discussions), it remains unclear to the reader of the reports whether beneficiaries and stakeholders are involved in the design phase of an evaluation (as implicitly requested by one of WVG's sub-criteria). However, to be fair on assessment, WVG may rethink to which extent participation of beneficiaries and 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 stakeholders were included during the interventions of a project. Thus, the fact that 16 out of 34 reports do not present clear evidence on this issue, discloses room for improvement. The same holds true for the inclusion of beneficiaries' perspectives on how to move on with the interventions after the end of WV's programmes, as 20 out of the 34 reports fail to address this issue appropriately. Despite this negative finding, the evaluation report from the ADP in Sierra Leone can show the way forward to improve on this sub-criterion as it provides a section on community engagement and deviates community-related recommendations.

This sober performance should, however, not be mistaken as overall failure of the evaluation teams to consider the sustainability of WV's interventions. In the majority of the reports the long-term perspectives of WV's interventions are discussed in a promising manner; therefore in this meta-evaluation 12 reports were rated as good and 18 as fair regarding this sub-criterion.

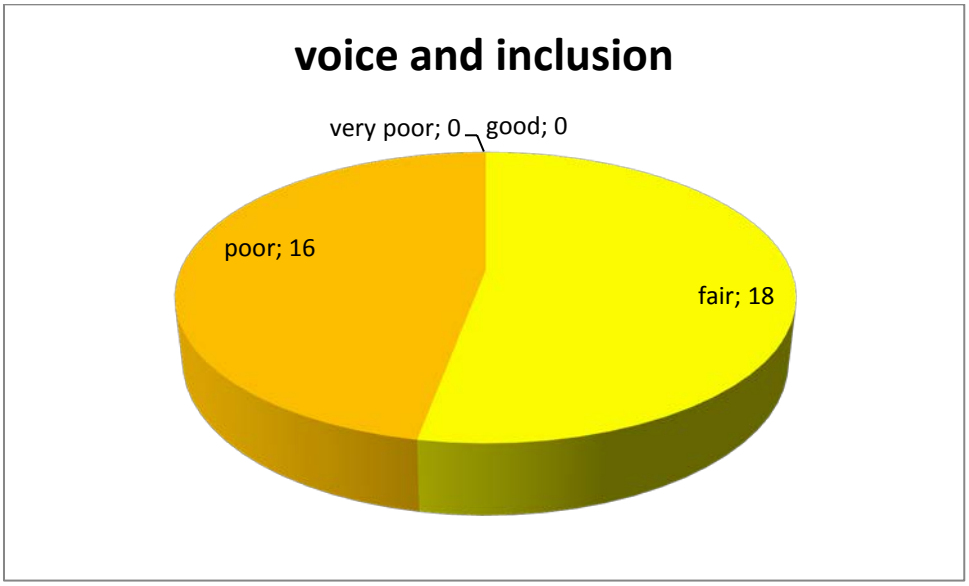


Figure 1: Overall performance referring to the criterion voice and inclusion

To sum up, figure 2 presents aggregated findings and reveals that overall 18 reports perform fairly good, while 16 are rather poor with regard to WVG's evaluation criterion "voice and inclusion".

4.2 Transparency

A sound 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 (11 rated poor, 3 very poor). Although 17 re-

ports describe and justify the size and the composition of their data sets very well (8 rated fair), 9 evaluation reports even fail to disclose the rationale behind their empirical base.

An additional weakness of 9 reports lays in the fact that authors objectivity cannot be ensured. Whereas some reports do not appropriately disclose information on its authors and their institutional affiliations, other reports were drafted by WV's staff and are therefore hardly free from self-consciousness. Beyond biases through the author, a broad body of literature has also shown that interviewer can bias results. Thus, it is a further point of critique that data for some evaluation reports was collected by WV staff.

Beyond these challenges of some reports, it is a good sign that 31 out of 34 reports are able to establish a logical link between the analysis conducted and their recommendations provided. Hence, despite room for improvement, figure 2 summarises that overall roughly 75% of the evaluation reports comply to a large extent with WVG's transparency criterion.

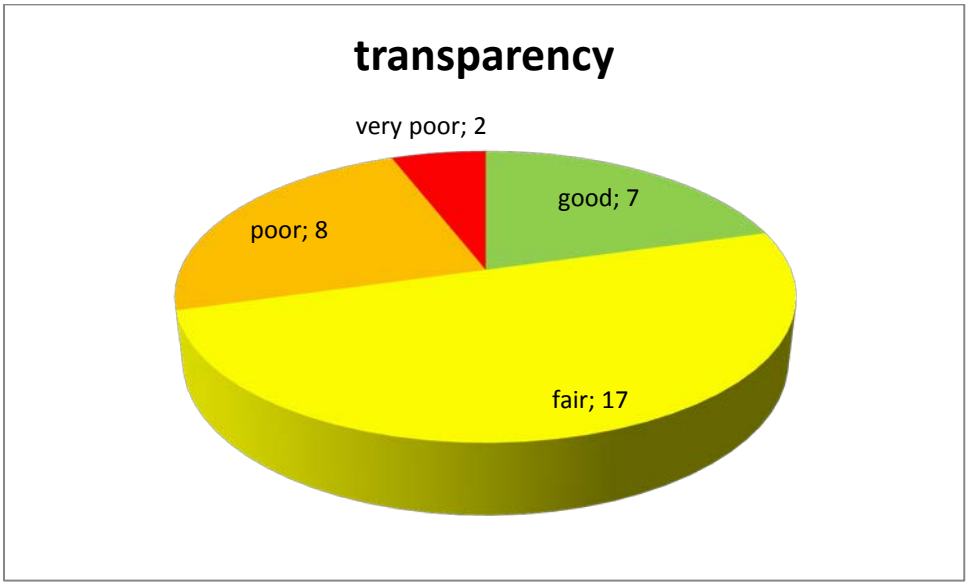


Figure 2: Overall performance referring to the criterion transparency

While half of the reports (17) perform fairly, 7 evaluations fulfil the criterion fully. Nonetheless, as 10 reports do not meet WVG's quality standards in a sufficient way, CEval suggests taking the evaluation report of the ADP in Burundi 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.

4.3 Appropriateness of evaluation 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 28 of the 34 reports state data collection methods which are relevant to the purpose of ADP evaluation (3 rated fair, 3 poor). This positive picture also holds for the appropriateness of collected data (20 rated good, 8 fair, 6 poor). This yields to a quite positive rating of overall appropriateness of methods (11 good, 16 fair) as displayed by figure 3.

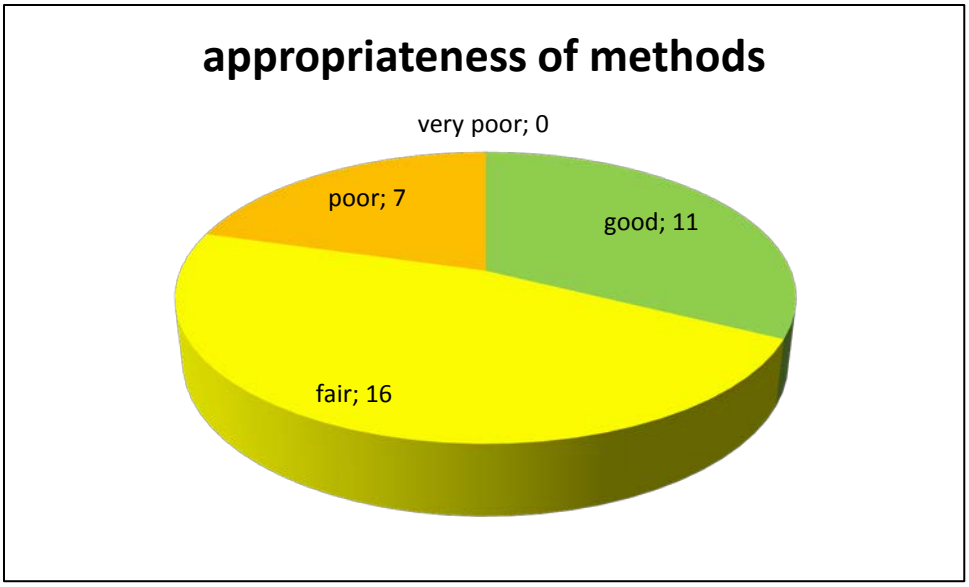


Figure 3: Overall performance referring to the criterion appropriateness

However, with regard to the validity of drawn conclusions, the quality of the evaluation reports deteriorates, thus roughly one third of the reports (11) derives inappropriate conclusions from a methodological point of view. Even though the remaining two thirds comply rather fairly 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 4.6).

4.4 Methodology

WVG's methodology criterion is quite far-reaching as it does not only focus on the articulation of the results chain and the underlying model of the evaluation, but also emphasise on data sources, data collection, and analysis methods. Thus, it is no surprise that given such wide-ranging aspects, the performance of the evaluation reports vary widely according to these sub-criteria. Whereas reports with a strong qualitative analysis agenda perform rather better with regard to introducing a results chain and programme theory (e.g. evaluation report of the Nong Son ADP in Vietnam), reports who centre rather around a quantitative analysis agenda are more successful in disclosing sampling methods (e.g. the report of the Yauli ADP in Peru).However, most evaluation reports display weaknesses in embedding their analysis into the results chain (28 out of 34 reports).

The picture for specifying sampling and analysis methodologies is much better as two thirds (23) of the reports cope adequately with this sub-criterion. Nevertheless, it has to be highlighted, that the remaining third (11) of the reports fails not only to critically appreciate the rationale for their quantitative and qualitative sample selection but also to discuss limits of the applied methodology.

Beyond these fundamental methodological concerns, it is further analysed to which extent the evaluation reports make reference to the capability and the robustness of the evaluated ADP's monitoring system. Unfortunately, the meta-evaluation discloses that the majority of reports (25) fail to address this aspect appropriately. The evaluation report of the ADP in Myanmar, however, marks a

positive exception as it refers to strength and weakness of the programme's monitoring and evaluation system and derives recommendation for future improvements.

At a glance figure 4 reveals a heterogeneous picture. Whereas 18 reports comply to a sufficient extent with WVG's methodology criterion (3 good, 15 fairly), 16 reports do not comply with this quality standard (13 poor, 3 very poor).

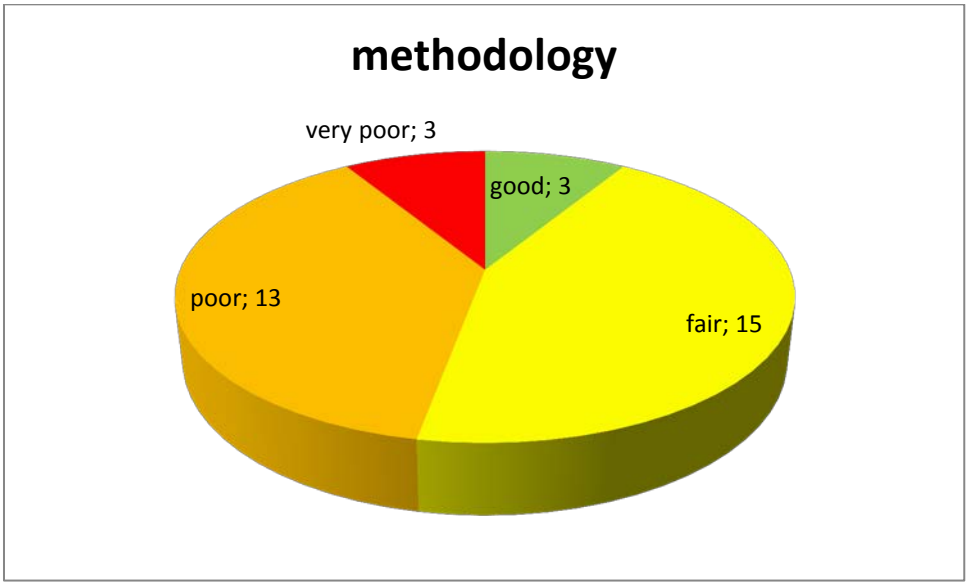


Figure 4: Overall performance referring to the criterion methodology

A detailed view at the reports discloses further, that evaluation teams often fail to particularly mention WV's quantitative data collection tools they applied or even worse that they lack the awareness of the wide range of WV's instruments and therefore did not apply them at all. As presented in table 1, only few reports refer to such instruments like for example the Functional Assessment of Literacy Tool (FLAT). Nevertheless, the vast majority of reports specify at least the use of standard quantitative methods like household surveys.

Table 1: Application of World Vision's data collection instruments

WV's data collection instruments	applied	not applied
Functional Assessment of Literacy Tool (FALT)	5	26 (3)
Development Asset Profile (DAP)	3	31
Caregiver survey	6	28
Youth Healthy behaviour	5	29
Measuring child growth	8	26

Note: Figure in parentheses indicates assessment of literacy by different instrument.
It is not controlled for applying instruments without reference to their origin.

Moreover, table 2 supports vast heterogeneity with regard to methodological variety. While most reports solely use qualitative standard instruments like focus group discussions or expert interviews, some reports attracted our attention due to the application of innovative qualitative methods like for example the tree of change, seed assessment or photo-voice. In particular, the evaluation report of the Nong Son ADP in Vietnam can serve as good example.

Table 2: Application of innovative qualitative methods

Innovative qualitative methods	applied	not applied
Comparison discussion group for non-sponsored children	4	30
Photo-voice	2	32
Seed assessment	5	29
Ladder of life	3	31
Tree of change	7	27

4.5 Triangulation

This section shows to which extent the ADP evaluations triangulate data, methods and perspectives in their analysis. According to figure 5, the overall picture for WV's triangulation criteria is quite similar to earlier findings on the methodology criterion. While half of the reports comply with WV's standards (2 rated good, 15 fair), the other half fails to do so (16 rated poor, 1 very poor).

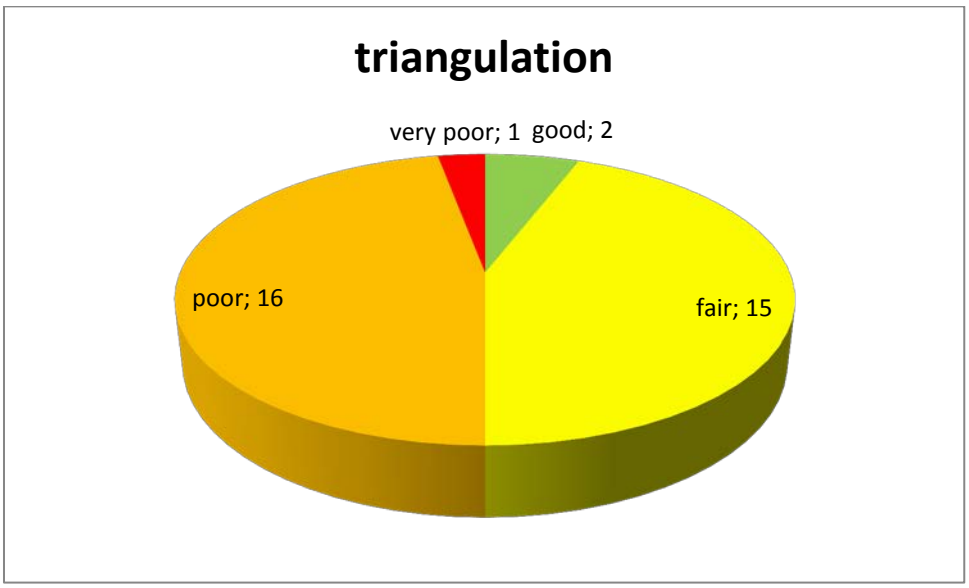


Figure 5: Overall performance referring to the criterion triangulation

A detailed look at the application of different data collection methodologies and data triangulation, however, causes greater optimism as 28 out of the 34 reports use a method mix based on a broad data base (2 rated fair, 4 rated poor). Unfortunately, this performance deteriorates when it comes to appreciating different perspectives of various stakeholders to explain how changes have occurred. Thus only 3 reports satisfy this sub-criterion in a good and 18 in a fair manner (13 rated poor).

Although the vast majority is on a state-of-the-art level regarding data collection, quality further decreases when it comes to appropriate discussion of results. Thus, 24 of the evaluation reports turn out rather poor in disclosing and interpreting apparently contradictory results (10 rated fair). Once again, given the broad heterogeneity among the reports, the evaluation report of the Bwembera ADP in Tanzania which performs well with regard to WV's triangulation criteria can be highlighted as a positive example with explicit statements on a mixed-methods approach to triangulate qualitative data from the tree of change method, key informant interviews, focus group discussions and site

visits with quantitative data emanating from the caregiver survey and a household survey as well as with insights from a literature review.

4.6 Identification of WV's contribution

This section illustrates to which extent the evaluation reports identify WV's contribution. A qualitative analysis on how the interventions of an ADP contributed to observed change is a good starting point. While nearly half of the reports tackle this issue (3 rated good, 13 fair), the other half fails to explore a clear link between ADP's actions, its outputs and following outcomes (18 rated poor).

To ascertain changes quantitatively it is important to use reference points. Baseline data allows observing differences between two points in time. Thus, it is rather unfortunate, that baseline data seems only available for 18 of the 34 ADP areas. By analysing collected data of project beneficiaries, data of a comparison group which was not affected by ADP's interventions is further important to isolate the impacts of an ADP from other factors (i.e. infrastructural improvements, natural disasters, governmental development programmes) which also yield to change in the project area. But again only some evaluation teams (6) make the attempt to exploit such data when deriving their conclusions.

Even worse, in 31 of the 34 reports reasonable (qualitative) considerations on such alternative factors are missing (26 rated very poor, 5 poor, 2 fair, 1 good). The fact that interventions of other actors could have also contributed to observed change is solely explicitly touch upon by the evaluation report of the Bwembera ADP in Tanzania. Finally, appropriate references to unintended or unexpected changes (negative or positive) are completely absent in the 34 evaluation reports.

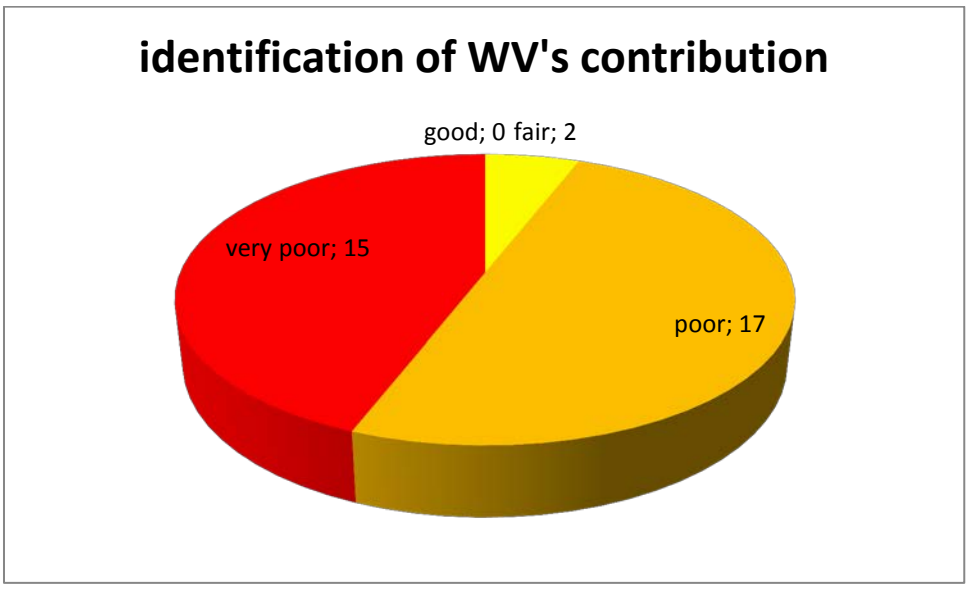


Figure 6: Overall performance referring to the criterion contribution

Hence, figure 6 reveals that overall the vast majority of the evaluation teams (17 rated poor, 15 very poor) did not appropriately comply with WV's contribution criterion as they did not manage to attribute observed changes to ADP's interventions.

4.7 Additional criteria

To assess the quality of the evaluation reports it is further checked to which extent they satisfied WVG's information needs. This section starts with a brief analysis whether the reports provide answers to the questions specified in the underlying terms of references (ToR). Moreover, the appropriateness of given recommendations and highlighted lessons learnt is considered.

The analysis reveals that many evaluation teams fulfilled the ToR in an acceptable manner; while some did not complete the assigned tasks (13 reports rated good, 13 fair, 7 poor, 1 very poor). A look at the recommendations shows that the vast majority of the reports (23 rated good, 8 fair), however, succeeds in providing appropriate proposals on how to continue with an ADP. Further, roughly two thirds of the evaluation teams went beyond deriving recommendations as in 22 of the reports lessons learnt during the evaluation are shared with WVG to allow improvement in future.

Overall assessment as displayed in figure 7 is thus predominantly promising. Although 12 reports cannot satisfy WVG's information needs (10 rated poor, 2 very poor), 15 reports provide a fair amount of information needed and the remaining 7 a good amount.

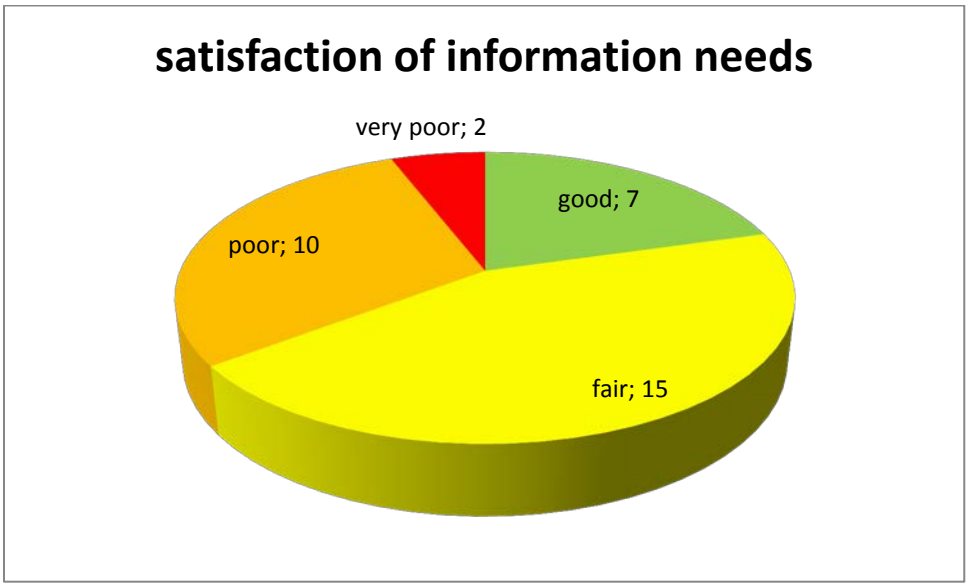


Figure 7: Overall performance referring to the criterion satisfaction of information needs

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. As display by figure 8 a great number of reports fail to do so (18 rated very poor, 2 poor); but given that 11 reports perform well in organising their findings (3 rated fair) the picture is quite heterogeneous. Thus, there are several reports which could serve as good example like the one from the Paucara ADP in Peru.

A view at figure 9 discloses that beyond organising findings, most evaluation reports completely fail to embed the analysis into a conceptual framework. Thus, 21 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 (7 rated good, 2 fair, 4 poor).

Figure 8: Performance referring to organisation of findings according to log frame indicators

In contrast, the importance of conceptualising change is exemplarily highlighted by the report on the Nong Son ADP in Vietnam, which allows the reader to understand cause-effect chains. Hence, regardless of particular evaluation results, it is of utmost importance that evaluation teams tackle this issue as ADP's outcomes can be only improved if one understands the structure that produces them.

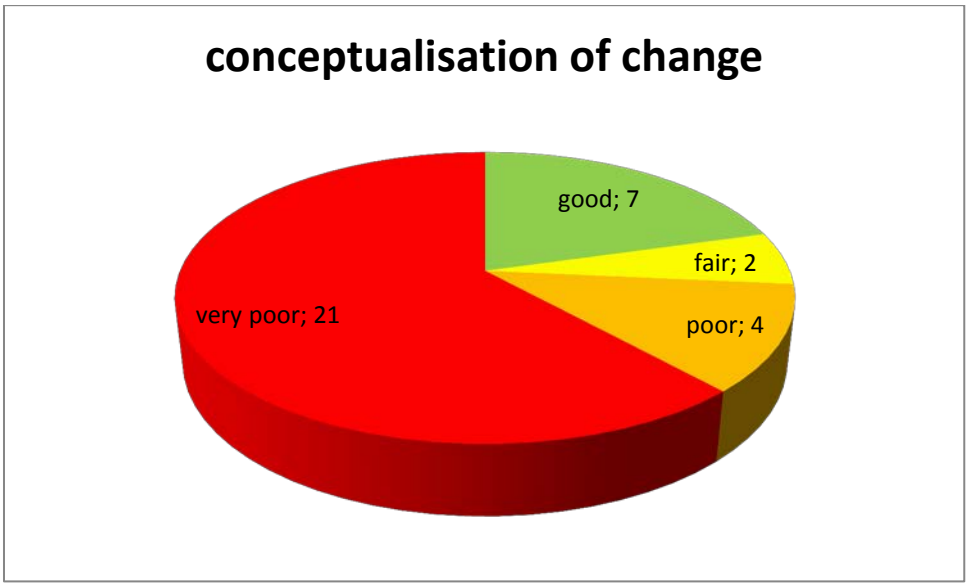
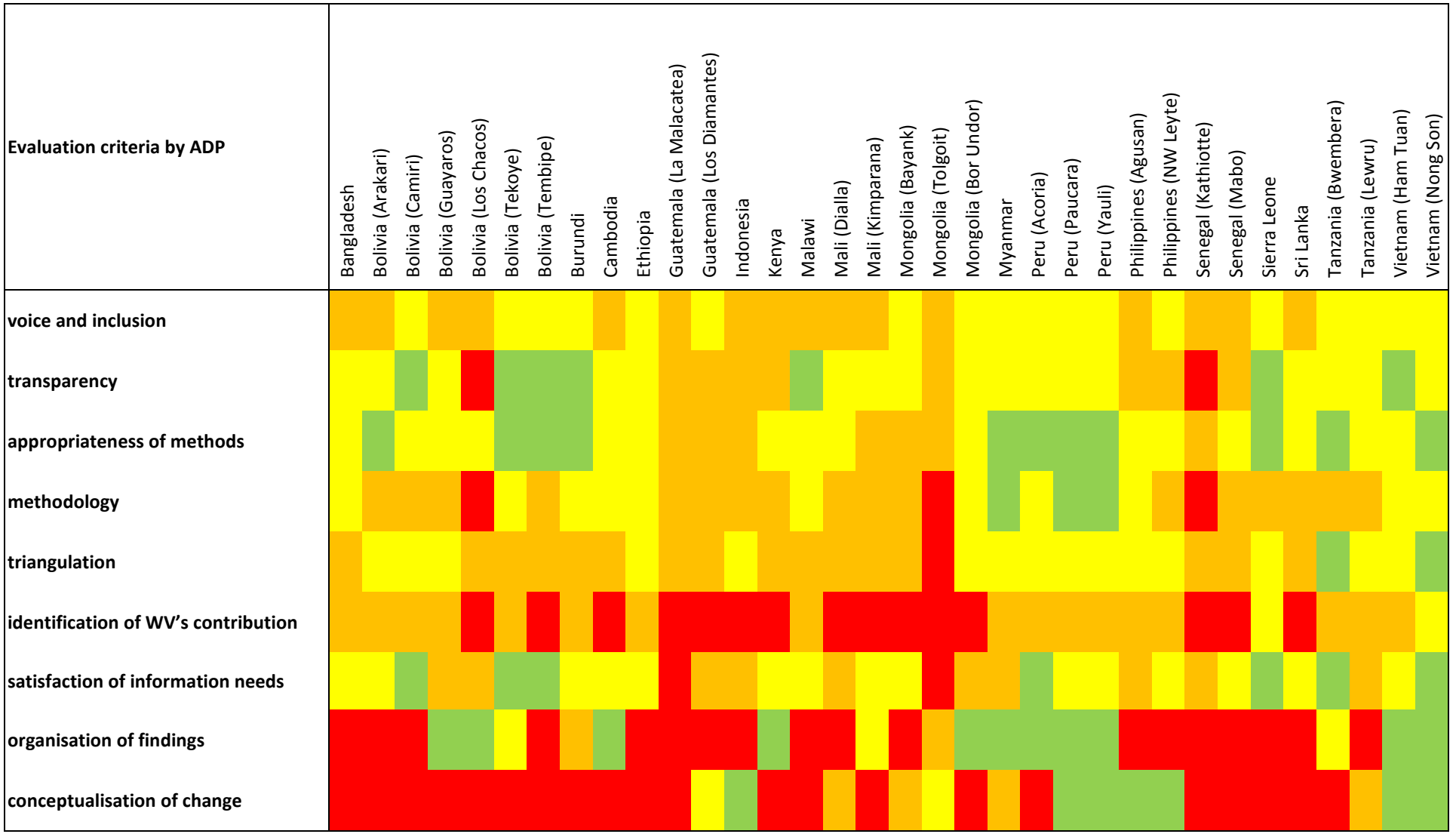


Figure 9 Performance referring to conceptualisation of change

5. Synthesis

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

Table 3: Evaluation results matrix



As summarised by the evaluation results matrix in table 3 many reports reveal serious shortcomings, while few cause optimism. Building regional clusters and analysing the reports in chronological order did, however, not disclose systematic differences. Thus, figure 10, a summary of the results of this meta-evaluation, highlights average strengths and weakness of the ADP evaluation reports best.

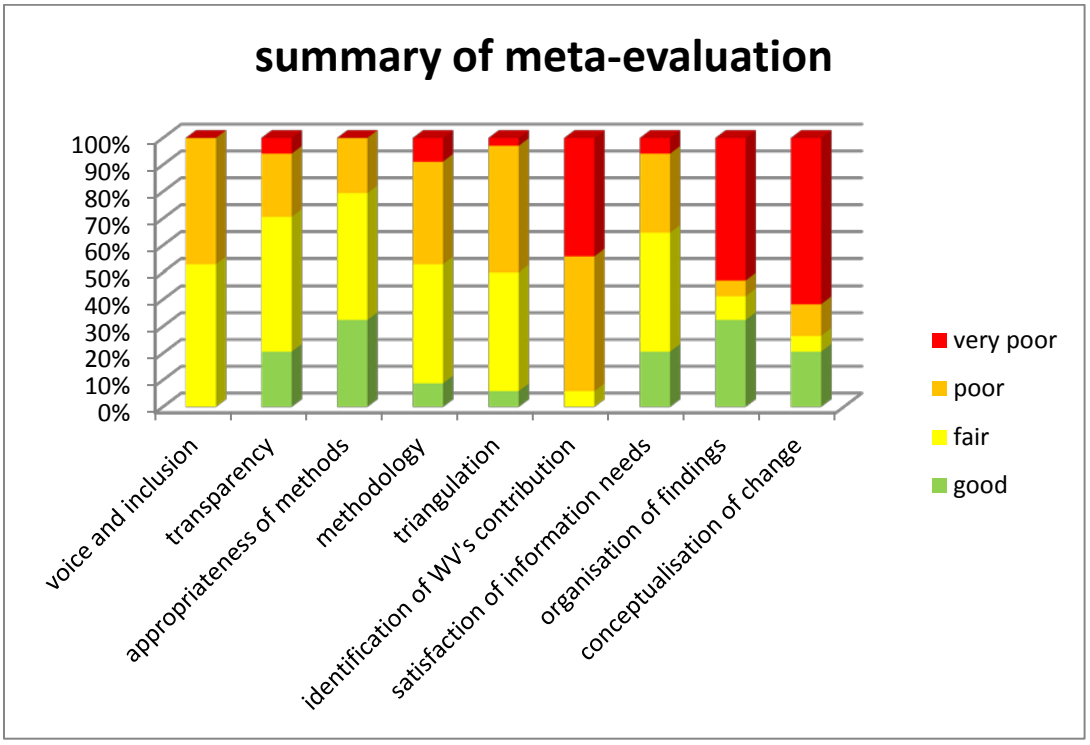


Figure 10: Summary of meta-evaluation results

While the reports on average were most promising with regard to WVG's transparency criterion and the appropriateness of applied methods (roughly 70% performed at least fairly), the criteria voice and inclusion, methodology, triangulation and satisfaction of information needs were at least in roughly half of the reports tackled in a sufficient manner (at least rated fairly).

In contrast, the fact that more than 60% of the reports did not present evaluation results in an easily understandable way against the programme's log frame indicators and even more than 70% refrained from conceptualising findings against a theory of change, raises serious concerns. The by vast alarming result of this meta-evaluation, however, is the deficiency of nearly all reports to detect WV's contribution to observed changes.

Given that several evaluation teams selected appropriate methods and applied them consistently, it is a pity that even they were unable to identify WV's contribution. The reason for this lies in the failure of the evaluation teams to differentiate between the gross outcome in the area under consideration and the net effects of an ADP's interventions. Whereas the net effects comprise positive and negative, intended and unintended changes caused by the programme, the gross outcome is the sum of (i) such net effects, (ii) extraneous confounding factors which are caused by other factors or actors (e.g. natural disasters, local government initiatives) and (iii) design effects (e.g. measurement errors) as displayed by figure 11.

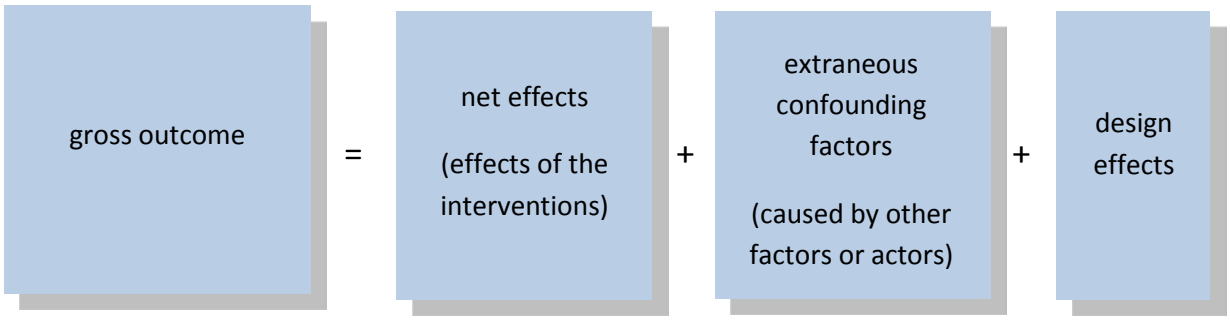


Figure 11: gross outcomes of an intervention

Thus, to derive valid conclusions on the effects of an ADP's interventions it is insufficient to focus solely on observed changes in an ADP area. Beyond identifying the gross outcome in an ADP area i.e. the overall change (Δt) measured between the start of the project (t_0) and the time of data collection for a particular evaluation (t_n), it is necessary to make reasonable considerations on its composition and on those effects which can be attributed to an ADPs' interventions. Hence, the main point of critique refers to the methodological fallacy inherent to evaluation reports claiming that an improvement of standardised indicators (e.g. an increase in school enrolment rates or a decrease in child mortality) can be directly interpreted as impact of an ADP's interventions.

6. Conclusion and recommendations

This meta-evaluation disclosed that WVG's independent ADP evaluations are not yet state-of-the-art. Whereas several important issues are often addressed in a promising manner, several gaps were identified. This calls for the following activities.

To improve the performance of its independent evaluations CEval recommends WVG to be even more precise on the design of ToRs as they already are. Thus, anchoring summarising tables which are organised along log frame indicators would enhance a rapid understanding of evaluation results and further facilitate comparability to evaluation results of other ADPs or in the long-run to earlier evaluation results of the same programme.

To prevent that underlying assumptions of the evaluation teams and ADP staff were not addressed in future evaluation reports, it would be further advisable to ensure that evaluation teams explicitly incorporate conceptualisation of change when developing the evaluation design and when drafting their reports.

To comply with the scientific quality criterion of objectivity it is recommended to contract independent evaluation teams. Thereby it is not only important that the lead consultant is not on the payroll of WV, but also enumerators or facilitators of focus group discussions have to be outsiders to the ADP. Otherwise, objectivity cannot be claimed convincingly.

Reliability is the extent to which an analysis yields the same result on repeated trials. Beyond objectivity it is an important scientific quality criterion. However, the extent of reliability can be only judged if the data generating process and analysis methods are clearly stated. As this is not the case for all evaluation reports of this meta-evaluation CEval advises WVG to demand in the ToR clarity on

those issues. Thus, evaluation teams should explain (i) why selected data collection methods are implemented, (ii) why the particular data collected is useful for the evaluation, (iii) why a particular set of methods is applied, and (iv) why each of this methods is appropriate, despite its inherent limitations. In conjunction with this set of questions, WVG could additionally stimulate the implementation of its data collection instruments when providing a link to the website or disclosing promising reports who could serve as best practice.

The third scientific quality criterion is validity. It refers to the extent to which an analysis really measures what it attempts to measure. As explained above, the majority of the evaluation reports do not yield to valid results as they fail to detach the net effects of and ADP's intervention from observed gross outcomes in the programme area. To enhance validity CEval recommends WV to reconsider their strategy and to think about pooling resources to evaluate in sum less ADPs but increase the validity of those programmes evaluated. Implementing quasi-experimental evaluation designs would be a feasible method to identify WV's contribution in a rather valid manner. Comparing baseline data and the data collected by the evaluation team from both, programme beneficiaries and a comparison group, would facilitate better attribution of an ADP's interventions as difference-in-difference designs control for extraneous confounding factors over time, and in turn allow isolating net effects for programme beneficiaries.

However, where quasi-experimental designs (or further advanced methods like randomised control trials) are not possible valid conclusions on the impact of an ADP cannot be derived from quantitative data. It is therefore even more important to point to limitations and to present descriptive statistics in a careful way by not only focussing on means, but also providing standard deviations. Moreover, in such cases, qualitative data could be helpful to at least give a hint on WV's contribution to observed changes.

The final recommendation refers to mainstreaming the ToRs with broadly agreed evaluation standards. Several ToRs specify already the OECD/DAC criteria relevance, effectiveness, and sustainability, but the two remaining criteria impact and efficiency are not mentioned. Although, they are partly implicitly tackled, highlighting them prominently to ensure assessment against this widely accepted set of standards is highly recommended. Again, it could be helpful to make in the ToR reference to OECD/DAC documents which could provide guidance to the evaluation teams from the outset. In addition the standards of the German Association for Evaluation (DeGEval) which comprises utility, feasibility, propriety and accuracy could be also mainstreamed in the ToR. For WVG this would have the particular advantage that its evaluations follow the same standards as one important donor of WVG, namely the German Ministry of Economic Co-operation and Development.