Michaela Raab and Wolfgang Stuppert

Review of evaluation approaches and methods for interventions related to violence against women and girls (VAWG)

June 2014
m.raab@posteo.de and wstuppert@gmail.com

Table of contents

Acknowledgments

This review has benefited enormously from the interest and support of our DFID counterparts, Clare McCrum and Zoe Stephenson, and the External Reference Group that has accompanied our work: Joëlle Barbot (CIDA), Krishna Belbase (UNICEF), Valeria Carou-Jones (UNFPA), Katie Chapman (DFID), Jennifer Leith (DFID), Helen Lindley (Womankind), Judith McFarlane, Jodi Nelson (Gates Foundation), Fiona Power (DFID), Amanda Sim (International Rescue Committee), Inga Sniukaite (UN Women), and Jeanne Ward.

Raja Litwinoff, Stella Maranga, Claudia Neymeyer, Jasmin Rocha and Vera Siber kindly pre-tested our interview guides and survey instruments. More than two hundred development and evaluation professionals have contributed to the review by making themselves available for interviews (list of interviewees in annex), participating in our survey on evaluation effects, providing evaluation reports or other literature, and facilitating contacts. Ian Askew (Population Council), Sabrina Evangelista (UN Women) and Leigh Stefanik (CARE) were particularly active in sourcing evaluations and getting us in touch with evaluation stakeholders. We are immensely grateful for the time and efforts contributed, given generously and without any material compensation.

Rick Davies and Carol Miller have accompanied our work with thoughtful comments on our blog www.evawreview.de and on Rick's blogs[1].

Last, but absolutely not least, we thank Julian Brückner (Humboldt University) who has provided technical advice on QCA at different stages of the review. We are indebted to Miruna Bucurescu, Scout Burghardt, Astrid Matten, Sanja Kruse and Paula Pustulka who made excellent use of their skills and knowledge in gender and social studies when coding thousands of pages of evaluation reports.

[1] The posts are available on http://mandenews.blogspot.de/2014/03/the-challenges-of-using-qca.html and http://mande.co.uk/2013/lists/me-blogs/a-review-of-evaluations-of-interventions-related-to-violence-against-women-and-girls-using-qca-and-process-tracing/

Executive summary

Review purpose and approach

This review ('the Review') was commissioned by the UK Department for International Development (DFID) to assess the strengths, weaknesses and appropriateness of approaches and methods used to evaluate interventions on violence against women and girls (VAWG).

A distinctive feature of the Review is the effort we have undertaken to understand **what makes an evaluation effective**. In our understanding, an evaluation is effective if it has an influence on programme implementation, policy and wider learning.

Methodology

The review team started with an analysis of 74 full evaluation reports of interventions focusing on VAWG. Subsequently, we conducted a web-based survey to assess four types of evaluation effects: (i) **action** effects (informing subsequent implementation of interventions on VAWG); (ii) **persuasion** effects (convincing others to support an intervention or the policies it advocates for); (iii) wider **learning** effects (influencing professionals beyond the evaluated intervention); and (iv) **empowering** effects on the intended beneficiaries of the intervention.

A sub-set of 39 evaluations was examined using **Qualitative Comparative Analysis** (QCA), to identify the *paths* (i.e. configurations of conditions) that led to effective evaluation in the field of VAWG. We found 28 out of 39 evaluations to have generated strong effects. Each effective evaluation is linked to at least one of eight identified *paths*.

Statistical analysis was used to assess trends and gaps regarding the methodology, quality and effects of the set of 39 evaluations.

For a more nuanced understanding of the interplay of conditions, we applied **Process Tracing** to five effective evaluations representing four different *paths*. Finally, to illustrate the range of methods used in VAWG-related evaluation, 13 **summaries of evaluations** were prepared (in annex).

Paths to effective evaluation

The Review has examined the following six ***conditions*** which we identified as contributing to evaluation effectiveness. Each condition is an aggregate of the factors listed after it:

- **Favourable context**: evaluation setting and complexity, and the evaluators' mandate.
- **Approach**: strong role of qualitative or quantitative methods in data collection, or a mix of both.
- **Compelling evidence**: compliance with established standards in research.
- **Sensitivity to the VAWG context**: sensitivity to evaluation-related risks and gender.
- **Participatory design**: active involvement of intervention stakeholders (implementers, donors, beneficiaries) in evaluation design and data analysis.
- **Good communication**: accessible presentation and wide dissemination of findings.

These conditions combine into eight distinct configurations or ***paths*** that QCA has found to lead to effectiveness in our set of 39 evaluations. **The choice of a particular approach or method was only one among several factors that combined with others to make an evaluation useful**. Purely qualitative, purely quantitative and mixed approaches were all found to generate useful evaluations, provided they followed one of the identified *paths*.

The first diagram below shows the paths that have led to strong effects, regardless of the evaluation context. The paths in the second diagram worked only where the overall context for the evaluation was favourable.

Evaluation commissioners can use these diagrams as decision trees to check whether an evaluation design reflects a combination of factors that has demonstrably led to strong effects in VAWG-related evaluations.

Paths to effective evaluation in any context

Path 1 (first line of boxes in the diagram below) covers some 54% of the effective evaluations in the set we have examined. It is composed of three necessary conditions: a strongly qualitative design, participation, and evaluators highly sensitive to the VAWG context. Whether the design includes quantitative methods or not is of no importance for this path.

In path 2 (covering 21%), participatory design and sensitivity play strong roles. It differs from path 1 only in that the evaluation approach needs to be strongly quantitative, i.e. evaluation findings are based chiefly on data gathered with quantitative methods.

In path 3 (covering 18% of the effective evaluations reviewed), participatory design, a strongly qualitative approach and good communication of the findings play major roles. Quantitative data, in turn, are absent or play only a minor role for the conclusions drawn in evaluations covered by this path.

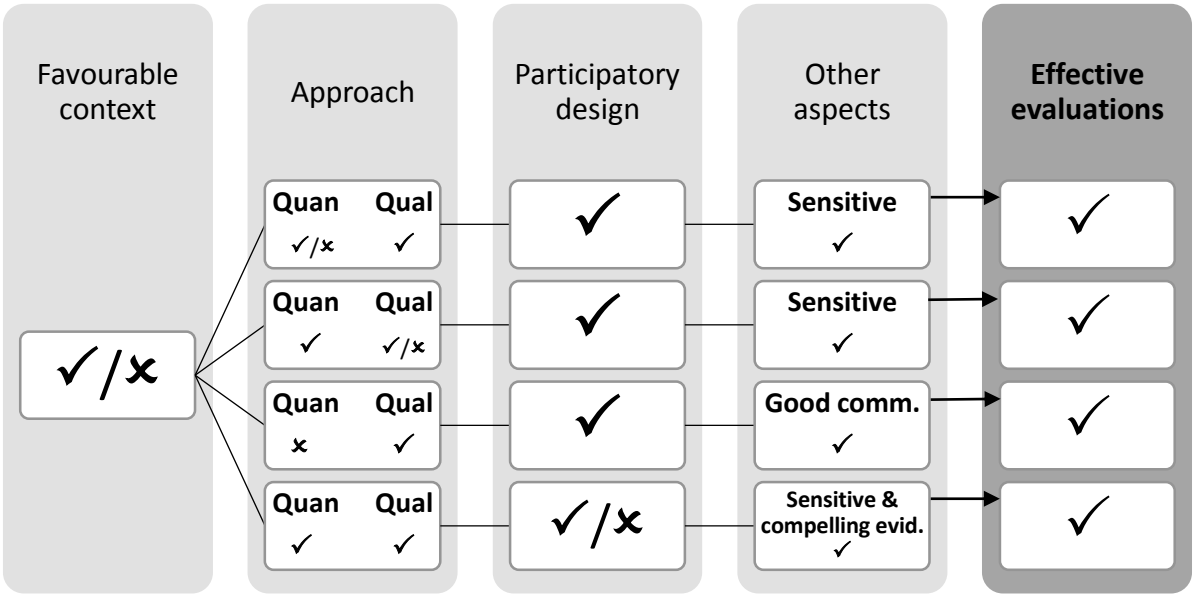
In path 4 (covering 11%), a strong mixed-methods approach, compelling evidence and evaluators' sensitivity to VAWG-related issues play major roles.

How to read the diagram

✓ only: The condition is necessarily present. In the 'approaches' column, ✓ means that the evaluation bases most of its conclusions on the particular type of data (qualitative or quantitative).

✗ only: The condition is necessarily absent. Under 'approaches', ✗ means that none or few of the conclusions are based on the respective type of data (i.e. the approach was not strongly qualitative/quantitative).

✓/✗ combined: The condition is not necessary; i.e. it does not matter whether the evaluation shows the characteristic or not. Under 'approaches', data of the respective type can be at the basis of none or almost all of an evaluation's conclusions (the approach may be strongly qualitative/quantitative or not).



Paths to effective evaluation in favourable contexts

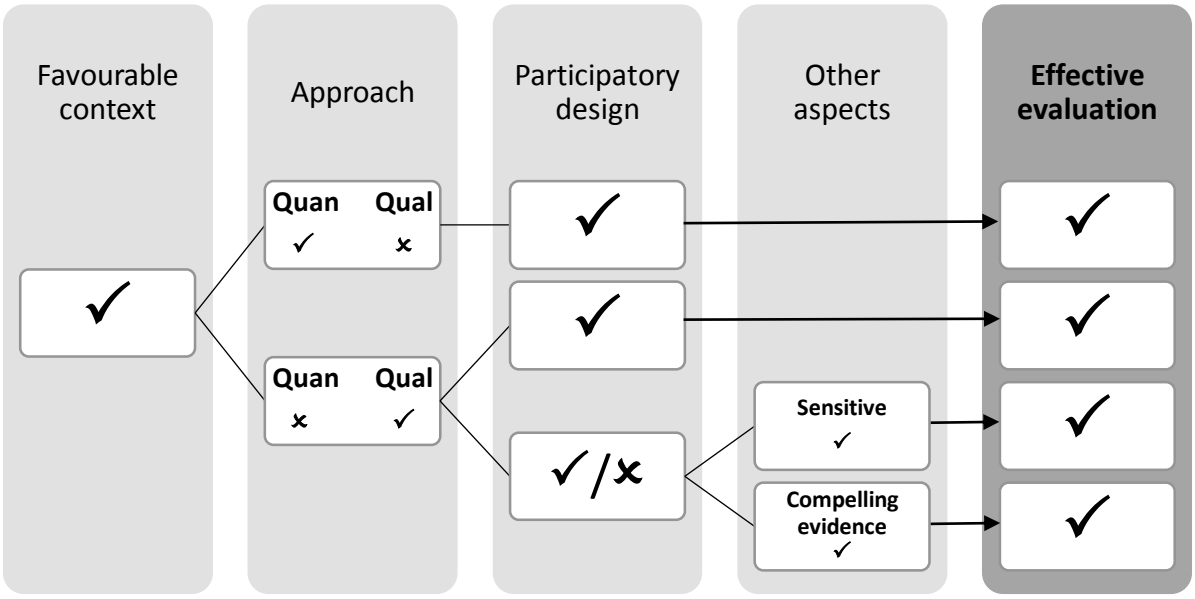
The *paths* below have led to strong positive evaluation effects in favourable contexts only.

In path 5 (top row, covering some 11% of effective evaluations reviewed), favourable context, participatory design and a strongly quantitative approach are important features of evaluations, whereas qualitative methods play only a minor role or no role at all.

Path 6 (covering 29%) differs from path 5 only in that the approach is strongly qualitative, and quantitative methods play no significant role or no role at all.

Path 7 (covering 36%) requires favourable context, a strongly qualitative design with no or only a minor role for quantitative methods, and an evaluation team sensitive to VAWG. A participatory design is not a requirement in this path.

Path 8 (covering 7%)[2] differs from path 7 only in that sensitivity to VAWG is not a necessary condition; instead, compelling evidence plays a major role.



What is a favourable context?

We defined 'favourable context' using three dimensions: (i) a stable internal and external environment for the intervention; (ii) relatively simple evaluation tasks; and (iii) evaluators with a strong mandate. We took the mean of these three dimensions to determine whether the context of an evaluation was favourable or not.[3]

[2] Percentages add up to more than 100 because one evaluation can be covered by several paths.

[3] For instance, where one dimension was rather present and two near-absent (0.66, 0.33 and 0.33), the mean was under 0.5, and the context was designated as non-favourable. The strong presence of one dimension could make up for the near-absence of two other dimensions, though. Where one dimension was strongly present and two other dimensions near-absent (1.00, 0.33 and 0.33), the mean was above 0.5 and we considered the context to be favourable.

Trends and gaps in evaluation quality

The 39 evaluations reviewed used **common data collection methods**: interviews, focus group discussions, surveys and desk review. 64% of the evaluations used mainly qualitative methods; only 2% were chiefly quantitative. 18% used mixed approaches with strong qualitative and strong quantitative aspects; 16% followed mixed approaches where either qualitative or quantitative data collection dominated.

Participatory design – identified in QCA as an important condition for effective evaluation – featured in 72% of the evaluations. Workshops near the start and the end of evaluations were used to ensure funders, implementers and beneficiaries of the intervention could contribute to the evaluation design and gain ownership of the findings.

Evaluators were rated as **knowledgeable about gender** studies and capable of discussing gender issues in virtually all the examined evaluations. However, 22% of the evaluations scored poorly on sensitivity to **evaluation-related risks: a disturbing finding** in view of the harm that VAWG research can cause to participants' rights and well-being.

The **quality of the evidence** produced in evaluation reports varied widely. ***Triangulation*** of data was often deficient: 44% of the evaluations based most conclusions on data from a single stakeholder group – most frequently, people implementing the intervention. In some 58% of the evaluations, we found ***bias*** to be likely in the selection of respondents. In 31% of the evaluations, power dynamics – for instance in focus group discussions – likely affected participants' responses. ***Patchy documentation*** of the evaluation terms of reference (TOR) and of data collection tools often made it difficult to gauge the validity of the findings.

Most evaluation reports were well structured and written in an accessible language. However, **communication** of evaluation findings and recommendations tended to be restricted to internal channels (report and verbal presentation to stakeholders).

Recommendations for evaluation practice in the field of VAWG

Methodological openness

Commissioners should be open to a wide range of approaches and methods, and encourage evaluators to tailor each evaluation to its specific purpose. Both qualitative and quantitative designs can lead to effective evaluation.

The diagrams on the previous pages can be used to check whether an evaluation reflects a combination of factors that has demonstrably led to strong evaluation effects in previous evaluations.

Methodological rigour

Compelling evidence has <u>not</u> emerged as the most important factor for effective evaluation – i.e., some influential evaluations may be based on weak data.

We assume that accurate data generates more well-grounded recommendations. Therefore, regardless of the methodology chosen, evaluation commissioners and quality assurance teams should engage in dialogue with evaluators to understand precisely how basic instruments of social research (surveys, interviews, focus group discussions and desk review) are used in the evaluation.

The approach and methodology should be transparently documented in the evaluation report. Annexes should include all relevant data collection tools, such as questionnaires and interview guides. Sampling strategies for surveys and/or for qualitative interviews should be spelled out.

Securing a favourable evaluation context

An unfavourable context limits the choices evaluators have. The choice of the right moment for the evaluation, appropriate resources, a strongly mandated evaluation team and a clear-cut evaluation task can make up for gaps in other context factors that we have measured.

Strengthening participation

Evaluations in the field of VAWG should be designed and interpreted in consultation with evaluation users (implementers of the intervention, donors and beneficiaries) to ensure evaluators obtain the right data, interpret it correctly and produce recommendations that are adapted to the evaluation purpose.

Sensitivity to gender and to evaluation-related risks

Evaluation teams need to be familiar with gender research, in particular in relation to VAWG. They must observe ethical guidelines, such as the WHO guidelines for research on violence against women and girls (WHO 2001), to prevent violations of the rights of those potentially affected by the evaluation.

Broader distribution for wider learning

Only one *path* for effective evaluation required good communication. Dissemination, however, is an important aspect of increasing the global knowledge base on VAWG. Evaluation reports should be published and distributed more widely, ideally in full. They should include full documentation of the methodology, and be shared via several channels including specialised list servers and social media.

Glossary of terms and abbreviations

QCA terms listed in this glossary are shown in *italic* typescript throughout this review report.

Approach	Set of data collection methods used in an evaluation.
Beneficiaries	All those receiving services (incl. training), goods or financial means as part of an intervention.
Calibration[4]	Process in which set membership scores are assigned to cases.
Complex solution term	Synonymous with conservative solution, i.e. a solution that is based solely on *configurations* of *conditions* that are deemed *sufficient* for the *outcome* based on empirical evidence.
Commissioner	The person who commissions an evaluation, i.e. who plans the evaluation and makes sure it happens (for instance by drafting the terms of reference, hiring the evaluators and making sure the report is published).
Condition	Factor which is used to explain the *outcome.* There are different types of conditions, such as *necessary, sufficient, SUIN* and *INUS* conditions.
Configuration	Combination of *conditions* which describes a group of empirically observed or hypothetical cases.
COVAW	Cost of Violence Against Women (name of an organisation).
Cases covered	The percentage of cases with membership in a given *path* of all cases that show the *outcome*.
Crisp set	Set which allows only for full membership (1) and full non-membership (0).
DFID	UK Department for International Development.
Equifinality	Allows for different, mutually non-exclusive *sufficient conditions,* or paths, for the outcome.
Evaluation	The systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation, and results in relation to specified evaluation criteria.
FGM/C	Female genital mutilation/ cutting.
Fuzzy set	Set which allows for partial membership, in addition to full membership and full non-membership. It enables the researcher to work with concepts for which the establishing of differences in degree among qualitatively similar cases is plausible and feasible.
GBV	Gender-based violence.
Intermediate solution term	Solution term based on easy counterfactuals. It is a superset to, and less complex than, the *complex* or *conservative solution term*.
Method	Individual data collection method.

[4] We have based the definitions of QCA terms on those provided in Schneider and Wagemann (2013).

Necessary condition	A *condition* can be interpreted as necessary if, across all cases, *set membership* in it is larger than or equal to each case's membership in the *outcome*.
Non-stakeholders	Individuals or groups of individuals that do not have a direct stake in a given intervention, but are important sources of contextual information for evaluators of the intervention. In the context of our review (and depending on the individual intervention), non-stakeholders may be country experts, thematic experts (gender, health etc.), NGOs implementing interventions similar to those evaluated or members of a community in which the evaluated intervention was not implemented.
Outcome	The phenomenon whose causes are studied in the QCA. [In the case of this review: positive evaluation effects.]
Participatory design	Evaluation design which allows stakeholders in the evaluated intervention and/or beneficiaries to influence the way methods are implemented and data interpreted.
Path	Logical combination of necessary *conditions* that is sufficient for the *outcome*.
Process Tracing	Case study method used to trace the process by which explanatory factors (independent variables, conditions…) lead to the explanandum (dependent variable, outcome…). Concatenating causal mechanisms is often central. These are often analysed on the basis of in-depth interviews.
QCA	Qualitative Comparative Analysis
Qualitative Comparative Analysis	The most formalized *set-theoretic method* which uses formal logic and Boolean algebra, and aims at establishing *necessary* or *sufficient conditions*, integrating parameters of fit.
Set-theoretic methods	Approaches to analysing social reality through the notion of sets and their relations. Can model causal complexity. QCA is one, but not the only set-theoretic method.
Set membership score	Numerical expression for the belonging of a case to a set. With *crisp sets*, only full membership and full non-membership are possible. With *fuzzy sets*, degrees of membership can be expressed.
Solution coverage	Percentage of all cases' *set membership* in the *outcome* covered by the *solution term*.
Solution formula/ term	The result of a *truth table* analysis. Usually consists of several *paths* (see also *equifinality*).
Stakeholders	In the context of our review, individuals or groups of individual who are in a contractual position to affect the design and implementation of an intervention, such as funding organisations, implementing organisations and governmental partners.
Sufficient condition	A *condition* can be interpreted as sufficient if, across all cases, *set membership* in it is smaller than or equal to each case's membership in the outcome.
TOR	Terms of reference (for an evaluation)
Truth table	At the core of any QCA, it contains the empirical evidence gathered by the researcher by sorting cases into one of the logically possible combinations, aka truth table rows. Each row linked to the *outcome* can be interpreted as a statement of *sufficiency*.
VAWG	Violence against women and girls

Review purpose and design

1. Purpose of the review

The purpose of this review (hereafter referred to as "the Review") is to generate a robust **understanding of the strengths, weaknesses and appropriateness of evaluation approaches and methods** in the field of development and humanitarian interventions on violence against women and girls (VAWG). It has been commissioned by the Evaluation Department of the UK Department for International Development (DFID), with the goal of engaging policy makers, programme staff, evaluators, evaluation commissioners and other evaluation users in reflecting on ways to **improve evaluations of VAWG programming**. Better evaluations are expected to contribute to more successful programme design and implementation.

2. Scope

The Review examines evaluations of interventions to prevent or reduce violence against women and girls within the contexts of development and humanitarian aid. The following criteria were applied when selecting evaluations for this review:

Inclusion criterion	Explanation
Evaluation	**Full evaluation report**[5] of an on-going or past intervention. OECD/DAC definition of evaluation: *"The systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation, and results in relation to specified evaluation criteria."*[6]
Evaluation context	International **development and humanitarian** interventions (including post-/conflict).
Types of interventions	Interventions **explicitly tackling any form of VAWG, as the main or major purpose** of the intervention. (For a typology of interventions, see Scoping Report).
Language	**English**.
Period	Evaluations **completed between 1 January 2008 and 31 December 2012**. An earlier starting date would have made it difficult to contact evaluation stakeholders for our survey and interviews. Evaluations completed after 31/12/2013 were excluded because too little time would have elapsed to assess their effects.
Publication status	**Published and unpublished** evaluations.

Table 1: Inclusion criteria for the reviewed evaluations

Compliance with specific quality standards was <u>not</u> part of our selection criteria, as we started out with the question: Which dimensions of evaluation practice make evaluations in this field effective? As a result, our initial set of evaluations comprised **all 74 evaluations** that met the criteria above. We identified these during the scoping phase using a combination of web search, DFID-related sources and snowballing via the internet and social media. Subsequently, the **set of evaluations analysed had to be reduced** as described in the Methodology section below.

[5] Summaries of evaluations, articles and meta-evaluations were not included, as they generally offered too little material for analysis of the evaluation methodology. Thus, some interesting evaluations had to be excluded because the documentation we obtained was limited to articles or summaries of evaluation findings.
[6] In keeping with the definition, we have included external and internal, mid-term and final evaluations.

The 74 evaluations cover a wide spectrum of interventions of varying complexity carried out by public and private not-for-profit actors all over the world, ranging from a single training project to complex multi-country programmes. Most evaluations occurred near or after the end of an intervention, with a smaller number of mid-term reviews.

Note: The larger the word, the greater the number of evaluations with the respective geographical coverage; 'global' refers to multi-country interventions across more than one region.

Figure 1: Geographical coverage

Other literature used in the Review has focused on four aspects: (i) evaluation quality, (ii) evaluation use and effectiveness, (iii) VAWG-specific issues in research and evaluation, (iv) concepts and definitions related to development and humanitarian work and its evaluation. This has included both published documents and grey literature, such as internal guidelines for evaluation commissioners.

3. Review methodology

Since the ultimate purpose of the Review is to improve evaluation effectiveness, we have examined both the characteristics of evaluations and the effects evaluations produce. The characteristics studied include, *inter alia,* conformity with established evaluation standards, methodological choices and circumstantial factors. Evaluation *effects* have been scrutinised both at the level of stakeholders in interventions on VAWG and their evaluation *(active stakeholders)*, and at the level of the intended beneficiaries[7] of the evaluated intervention.

We believe that no single evaluation design generates optimal effects under all circumstances. A specific kind of evaluation may be the best choice in some situations, but a deficient one in others. The analytical method that forms the backbone of this review,

[7] We realise the term "beneficiaries" symbolically assigns a passive role to those whose lives are expected to improve as a result of an intervention. We consciously continue to use the term to operate a clear distinction between those who are in a position to shape the course of an intervention and those who have a less direct influence.

Qualitative Comparative Analysis (QCA) allows the identification of different *paths*, i.e. configurations of factors that produce the desired effects (*outcomes* in QCA terms).

Language and form: QCA uses relatively simple terms, but is complicated in terms of its logic. In view of the purpose of the Review, we have opted for a concise report with relatively little technical detail. Additional technical information is available in the annexes, as well as in the separate Scoping and Inception Reports. A more scholarly article on our methodology has been prepared for publication in a peer-reviewed journal.

Terms in *italic* typescript are part of the glossary on the first pages of this report.

3.1 Qualitative Comparative Analysis (QCA)

QCA is based on the assumption that several cause-to-effect chains coexist for identical effects. It examines sets of ***conditions*** in relation to specific ***outcomes***. In our Review, the approaches and methods, the context in which evaluations take place and the adherence to quality standards, are *conditions.* Evaluation effects are the *outcome* that we have studied.

QCA helps identifying ***paths***, i.e. combinations of conditions that are sufficient to produce the *outcome*. The group of *paths* identified for a set of cases is called a ***solution*** in QCA. In this Review, the *solution* describes all combinations of evaluation characteristics that lead to effectiveness among the evaluations analysed.

Reliability

The specifications of our QCA satisfy the recommendations made by experts in the field, such as the minimum number of cases required for the number of conditions in our analysis. Our findings are therefore reliable with regard to the 39 evaluations that we have analysed.

Comparison of our set with the total of the evaluations obtained in the scoping phase suggests that we covered all important variants of existing evaluations. Hence, as long as important parameters of evaluation practice stay within known boundaries, our findings provide meaningful guidance for the assessment of future evaluations in the field.

Further reading: Annex II (Methodological notes) includes additional information on QCA.[8]

3.2 Main steps of the Review

3.2.1 Scoping the evaluation landscape

We used a three-pronged search strategy, combining (i) web search, (ii) communication with our contacts in the fields of VAWG and evaluation, and (iii) snowballing via these contacts and specialised list servers. The Scoping Report describes this process in detail.

3.2.2 Assessing evaluation characteristics

Initial coding of 74 reports

Preliminary coding examined the initial set of 74 evaluation reports to assess which data was available across the set, and what quality standards the evaluations fulfilled. The reports proved to be highly diverse in terms of form and size (8-258 pages, median: 51 pages). Data on context and resources for the evaluation was rare across the set. Subsequently, we included questions on those aspects in the survey with evaluation stakeholders (see 3.2.3 below).

[8] For QCA as a scientific method, see Ragin 1987 and 2008, and Schneider and Wagemann 2012. QCA in evaluation: Befani 2007 and 2013.

Identifying evaluation stakeholders and preparing the survey

We identified evaluation stakeholders via (i) persons who had forwarded evaluation reports to us, (ii) the websites of the organisations involved, (iii) snowballing through these contacts and (iv) an appeal for support via the Review blog.

3.2.3 Learning about evaluation effects and narrowing the field of enquiry

Preparation of the survey

A total of 13 persons representing different perspectives – evaluators, commissioners, implementers of interventions, funders – agreed to be interviewed on evaluation effects. Their input, as well as feedback from the Review Reference Group on our initial definitions, informed the design of our web-based survey on evaluation effects. The survey was pre-tested by 5 international professionals with expertise in VAWG-related interventions (in implementing, funding and evaluating roles).

Survey implementation and response

The survey was designed using the open source programme LIME Survey, asking different sets of questions depending on the respondent's role in the evaluation analysed.

We contacted 212 evaluation stakeholders, striving to reach - for each evaluation - a commissioner, an evaluator and a representative of an organisation that had implemented the evaluated intervention. Extensive e-mail correspondence helped to obtain an excellent response rate of 70%. Response data, initially hosted on the commercial LIME server, was transferred into a common statistics programme (SPSS) for analysis.

Narrowing the field of enquiry

Upon completion of the survey, we retained only those 39 evaluations for which we could obtain survey data from at least two different perspectives – an evaluator's and that of a representative from the implementing organisation.

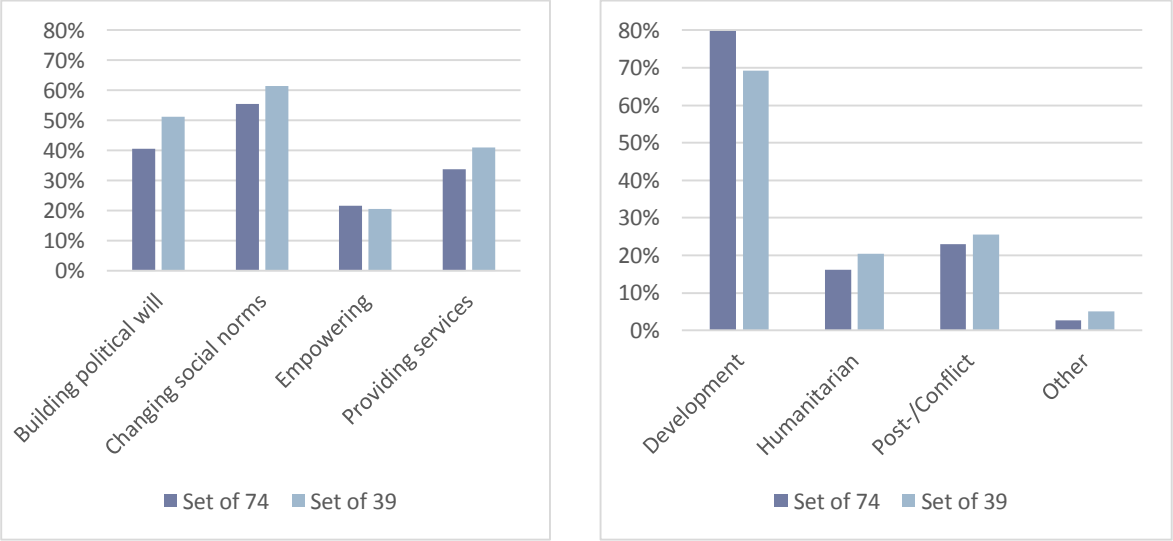


Figure 2: Themes & contexts of the sets

As shown in figure 2 above, we did not detect any significant differences between the initial set of 74 and the set of 39 evaluations in terms of themes and broader context.

QCA was performed on the set of 39 evaluations. Subsequently, 5 effective evaluations representing different combinations of methods were examined through Process Tracing.

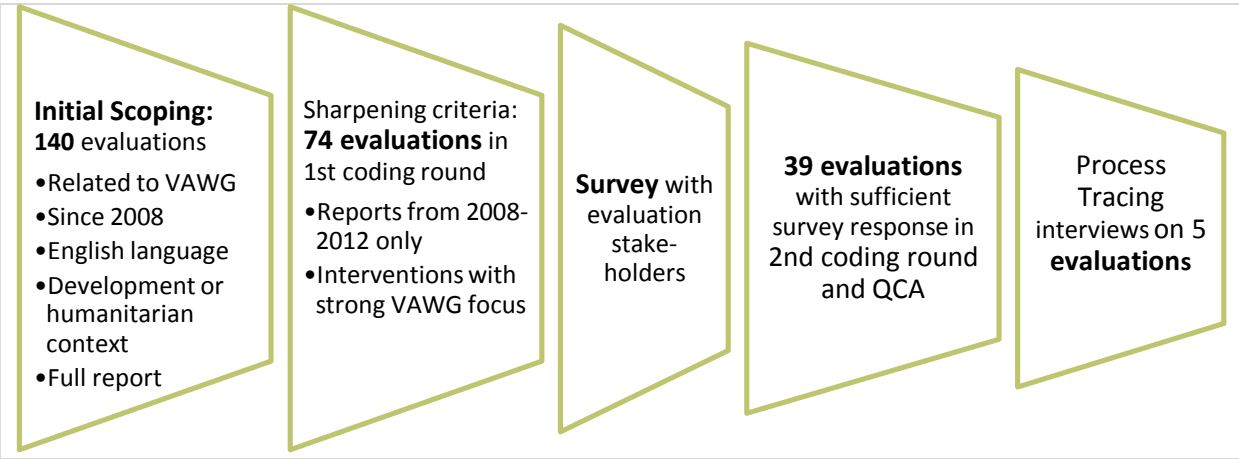


Figure 3: Narrowing the field of inquiry

Calibrating conditions

A second coding round was performed on the QCA set of 39 evaluations, to obtain specific data on the *conditions* for positive evaluation effects, and to examine aspects of evaluation quality in greater depth. We defined *conditions* to capture the evaluation context and quality standards (including quality of evidence, participatory design, sensitivity to GBV context, and communication of findings), and introduced two *conditions* representing broad approaches – qualitative and quantitative data collection methods.

For each condition, we devised rules to determine the extent to which it was met, for instance at what point the design was strongly participatory, or when we rated an approach as strongly qualitative, strongly quantitative or both. Subsequently, we analysed the 39 evaluations and assigned scores reflecting the degree to which each condition was met by each evaluation.

Further reading: More detailed information on all processes outlined in 3.2.1 – 3.2.3 above is provided in the separate Scoping and Inception Reports. Data collection instruments are included in annex to this report.

3.2.4 Reviewing evaluation approaches and methods

Using the open source programme *fsQCA,* we carried out an initial Qualitative Comparative Analysis with *conditions* capturing aspects of the evaluation context, evaluation quality, and the methodological approach. QCA showed that some configurations of *conditions* simultaneously resulted in effective and ineffective evaluations. We re-examined our data to identify the factors that accounted for those contradictions, and redefined the *conditions* accordingly. We arrived at a set of seven *conditions* covering evaluation context, evaluation quality (four *conditions*) and methodological approach (two *conditions*).

Subsequently, we used QCA to arrive at a *complex* and an *intermediate solution,* each with several *paths*. The *intermediate solution* is based on the assumption that the presence of a favourable context and the fulfilment of each of the four aspects of evaluation quality contributed to positive evaluation effects. Finally, we analysed the *paths* present in the intermediate solution and their implications for the choice of evaluation design.

3.2.5 Tracing success and describing evaluations

Process Tracing

Upon completion of the QCA, we examined five evaluations representing four different *paths*, using **Process Tracing** to explore links between *conditions* and *outcomes*. Three of the *paths* apply to the largest share of evaluations for different types of approaches; the fourth *path* has the highest overall coverage. For each of these *paths*, we chose typical cases with strong effects. We conducted semi-structured phone interviews with 2-4 stakeholders per evaluation (12 interviews in total) to find out how the *conditions* in each *path* played out. The interviews were transcribed and coded.

Descriptions of evaluations

Part of this assignment was to produce short **descriptions** of evaluations that presented suitable approaches and methods. Our selection of evaluations for the descriptions was primarily guided by the intention to show a wide spectrum of approaches and methods – not by findings from the QCA. Most evaluations were from the set of 39 (as the richest data was available on that set); one was drawn from the initial set of 74 (Jackson 2012). We added the influential IMAGE study (Watts *et al.*)[9] to include a randomised controlled design.

3.3 Difficulties and limitations

We followed high standards of rigour throughout the Review, a six-month desk-based exercise for scoping, data collection and analysis.

Limited data availability

The evaluation reports were our main source of data. Primary data collection – our survey with evaluation stakeholders and interviews – was constrained by the availability of our survey participants, who could not be expected to dedicate more than 15-20 minutes to the survey. As a result, we kept the number of survey questions to a minimum.

Limitations in the measurement of certain factors

- Persuasion effect: We used (i) the widening of intervention stakeholders' networks as a proxy for advocacy success, and (ii) the evaluation causing continued or additional funding to the intervention as a proxy for accountability-related positive effects.
- Learning effects: It was difficult to assess whether the findings of 39 evaluation reports generated learning beyond the evaluation stakeholders. As proxy measures, we assessed the potential for learning effects by examining the publication status (unpublished evaluations are unlikely to influence outsiders), the presence or absence of media reports on the evaluation, and whether any evaluation findings were surprising to the users.
- Long-term effects couldn't be measured. To comparatively assess the effects, every evaluation is required to have the same chance of causing such effects. The most recent (dating from 2012) had only one year for its effects to develop.
- Cultural sensitivity: A technical problem caused by the LIME Server[10] made that no survey responses on cultural sensitivity were recorded. This dimension of sensitivity was subsequently removed from our analysis.

[9] The study was published before 2008, which made it ineligible for this Review.

[10] We ran a test with specialised software to pre-test data processing on LIME but for unknown reasons the bug in LIME remained undetected. We have reported the issue to LIME.

Reliability of data

The **survey respondents** were individuals with stakes in the evaluations. We provided anonymity to encourage authentic answers. The diversity of responses obtained, including some stark criticism, indicates that respondents felt able to voice their opinions freely.

Inter-coder reliability was achieved with detailed instructions and reporting forms, and by swapping evaluation reports between coders so that each report was coded by at least two different coders.

Limits in nuance

Evaluations come with many characteristics and happen in highly diverse contexts. One could envisage a study of dozens of factors that contribute to effective evaluations. However, the number of *conditions* examined in QCA is limited by the number of cases that enter the analysis. That is why we have used seven broad *conditions* for QCA (for instance, 'compelling evidence') rather than a larger number of individual factors.

In our QCA, we examined a set of 39 evaluations. We aggregated the characteristics of evaluations and their contexts into **seven *conditions*** likely to influence evaluation *effects* (see section "*Conditions for evaluation effectiveness"* below). As each *condition* could be present or absent, 128 configurations were possible (2 to the power of 7). Of those configurations, 23 were covered by actual cases in our set. That coverage was appropriate to anchor our findings in empirical evidence.

To make full use of the data we have gathered, this Review provides – in addition to the QCA findings - analyses of specific evaluation characteristics. We have used statistical tools to analyse trends and gaps in the set of 39 reviewed evaluations. Process Tracing was used to create more detailed analyses of five exemplary cases.

Review findings

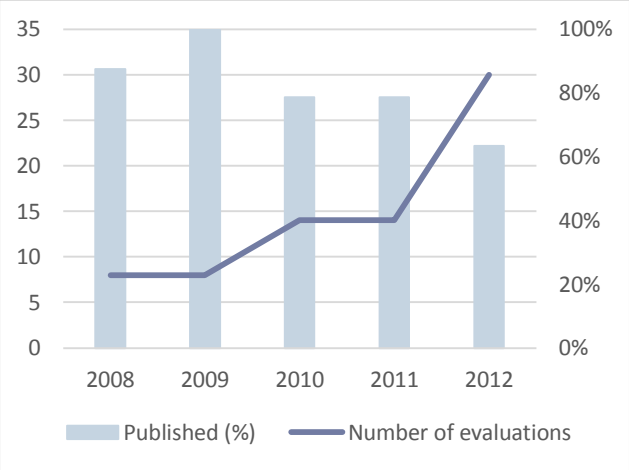
4. Trends and gaps in VAWG-related evaluations

This chapter starts with information on the characteristics of VAWG-related evaluations in development, humanitarian and (post-) conflict contexts as observed in our initial set of 74 evaluations. Sections 4.2-4.4 focus on the QCA set of 39 evaluations. We provide definitions of the evaluation characteristics, as well as descriptive statistics on (i) evaluation approaches and methods, (ii) gaps in the compliance with quality standards, and (iii) the contexts of the evaluations. The last section describes the evaluation effects we have found.

4.1 Number of evaluations

Overall, the number of evaluations in the field of VAWG appears to have increased. However, many evaluations were unpublished for various reasons, including the need for secrecy in particularly sensitive contexts.

The evaluations found in the scoping phase covered all DFID thematic priorities (as per the DFID theory of change: empowering women and girls, changing social norms, building political will and institutional capacity, and providing comprehensive services) and intervention contexts (development, humanitarian and conflict-related). No significant trends over time or gaps were observed regarding themes and contexts.



Data from the 74 evaluation reports coded during the scoping phase of the Review seemed to indicate a decrease in the percentage of published evaluation reports.

That does not necessarily mean a downward trend in the publication of evaluation findings. Our graph shows only evaluation reports that have been published in full. That is, it omits evaluation summaries and articles presenting the evaluation findings.

Fig.4: Number of evaluations (set of 74 reports)

4.2 Approaches and methods

This section focuses on methodological choice. Evaluation quality and the quality of the implementation of the methods are discussed in section 4.4.

Key definitions: For precise measurement, we have defined some terms in a very specific manner. Some of these definitions are narrower than others that are commonly used in evaluation.

We define an **approach** as the set of data collection methods used in an evaluation, while the term **method** refers to an individual data collection method.

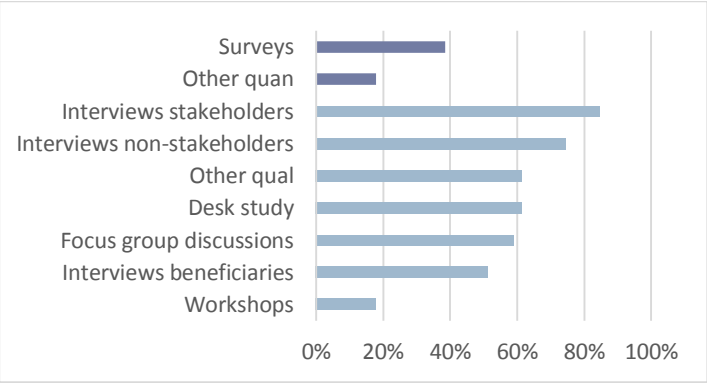
Participatory design refers to an evaluation design which enables donors, implementers and/or beneficiaries to exert an influence on the way evaluation methods are chosen and implemented, and data interpreted.

We identify an approach as **quantitative or qualitative** depending on the methods used to collect the data underpinning the evaluation conclusions. Data collection methods that require the researcher to pre-determine possible answers (for instance in standardised survey questionnaires) are categorised as quantitative (**quan**). If data collection is open to unsolicited information (for example in semi-structured interviews), we deem the method to be qualitative (**qual**).

Finally, we have assessed gender sensitivity by asking questions about the evaluators' familiarity with gender studies and their ability to produce a nuanced discussion of gender in the evaluation report. Data had to be collected and analysed with gender differences in mind.

4.2.1 Methodological choice

All 39 evaluations in the QCA set relied on common qualitative and/or quantitative data collection methods, with qualitative methods being most frequent. Some 85% of the evaluations included qualitative interviews with intervention stakeholders – usually representatives of the implementing organisation. Interviews with non-stakeholders (for instance, specialists external to the intervention) and desk studies came second (with 74%) and third (with 62%). Surveys were used in 39% of the cases, and other quantitative methods in 20% of the evaluations reviewed.



"Other quantitative methods" included *inter alia* health facility assessments using pre-designed checklists and inventories.

"Other qualitative methods" included field observation, safety audits and social influence maps.

Figure 5: Data collection methods

We qualified a design as purely quantitative when the evaluator only used quantitative methods, or when qualitative tools played no apparent role for the evaluation conclusions. "Quantitative design with qualitative backing" meant that most conclusions of the evaluation were grounded in data obtained through quantitative data collection tools alone; only some conclusions were based on qualitative data as well.

"Mixed approach" applied only to evaluations where most conclusions were backed both by data obtained through quantitative methods and data from qualitative methods.

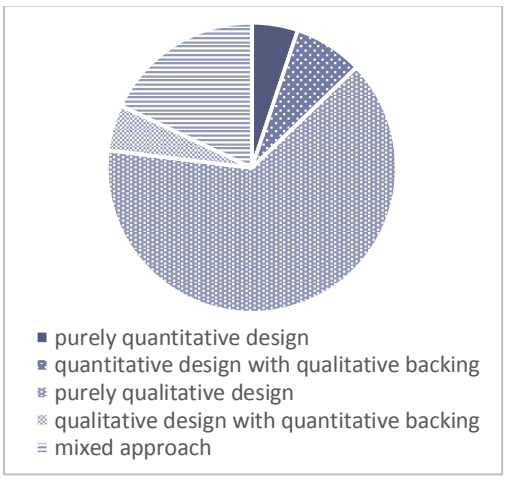


Figure 6: Evaluation design (set of 39 evaluations)

4.2.2 Participatory design

Most evaluators chose a participatory design, i.e. workshops were held near the beginning of the evaluation to discuss and fine-tune evaluation design, and near the end to reflect on preliminary findings. In 72% of the evaluations at least two such workshops were held with stakeholders in the intervention.

Our interviewees emphasised the importance of participation as a way to generate trust and strengthen ownership among evaluation stakeholders. As one evaluation commissioner stated: "*to the degree possible we do engage and discuss and share ideas with policy makers and program managers to make sure that the information is relevant and useful. […] The programme managers and the policy makers, they are much more vested when you engage them right from the beginning of the process, rather than bringing them in at the end"* (Kim 2009 #2). Participation also tended to deepen the evaluators' understanding of the intervention and its participants: *"I think there is a kind of paradigm that research must be something that should be kind of kept at arm's length in order to be objective but […] there is [another] way to conduct research that is rigorous and objective. […] to some extent researchers actually do need to get in there and understand the situation in some way in order to really understand what it is that they are studying"* (Kim 2009 #3).

Degree of participation	
No participation (as defined in the Review)	2,6%
At least one workshop was held in which evaluation stakeholders were able to either discuss the design of the evaluation or preliminary results.	25,6%
At least two workshops were held in which evaluation stakeholders were able to discuss both the design of the evaluation as well as preliminary results.	28,2%
At least two workshops were held in which evaluation stakeholders were able to discuss both the design of the evaluation as well as preliminary results. Ultimate beneficiaries participated in at least one of them.	43,6%

Table 2: Degree of participation in the reviewed evaluations

4.3 Evaluation quality

We examined quality in terms of the extent to which the evaluations in the set of 39 fulfilled established standards with regard to (i) the evidence they presented, (ii) ethical aspects of the research process, (iii) gender sensitivity, and (iv) presentation and distribution of evaluation findings, conclusions and recommendations.

4.3.1 Sensitivity to gender and ethical issues

All evaluations in our set assessed interventions intended to reduce violence against women and girls. Violence against women and girls is connected to power, aggression and potentially traumatising experiences.

Gender sensitivity

Evaluations of interventions against VAWG are part of the broader field of gender research. We measured gender-sensitivity by asking questions about the evaluators' familiarity with gender studies and their ability to produce a nuanced discussion of gender in the evaluation report. Data had to be collected and analysed with gender differences in mind. Such gender sensitivity was found to be strong for the vast majority of the examined evaluations.

Aspects of Sensitivity	Absent	Weak	Strong	Very strong
Gender sensitivity	2,6%	0,0%	48,7%	48,7%
Sensitivity to evaluation-related risks	2,7%	18,9%	21,6%	56,8%

Table 3: Sensitivity to gender and to evaluation-related risks

Sensitivity to evaluation-related risks

Evaluations in the field of VAWG come with physical and psychological risks for the persons involved. We considered evaluation teams sensitive to such risks, if they took precautions to respect the rights of informants, in particular VAWG survivors, so as to prevent any harm potentially caused by the evaluation process or its publication.

In 78% of the evaluations, strong or very strong sensitivity to evaluation-related risks was reported. However, almost one-quarter of evaluators appeared largely unaware of serious risks to the rights and well-being of those involved in the evaluation.

In view of the limited information available in most evaluation reports, we were not able to verify whether all evaluations systematically respected **ethical guidelines**. As a proxy, we measured the evaluator's awareness of risks. Anecdotal evidence suggested that practices varied – from rigorous ethical review processes to more casual approaches, which potentially caused safety risks and human rights issues.

4.3.2 Quality of evidence

We examined the extent to which evaluations complied with established research standards, in particular the use of original data (instead of exclusive reliance on secondary sources), the prevention of bias, and data triangulation.

Some two-thirds of the 39 evaluations commendably based their conclusions on data gathered as part of the evaluation (original data). However, major gaps in the quality of data collection were common.

In 51% of the 39 evaluations we examined, two of the three dimensions of quality of evidence (as explained in the paragraphs below) were present.

In 28.2%, all three dimensions were present.

In 5.1%, we did not find sufficient evidence for any of the three dimensions of high quality of evidence.

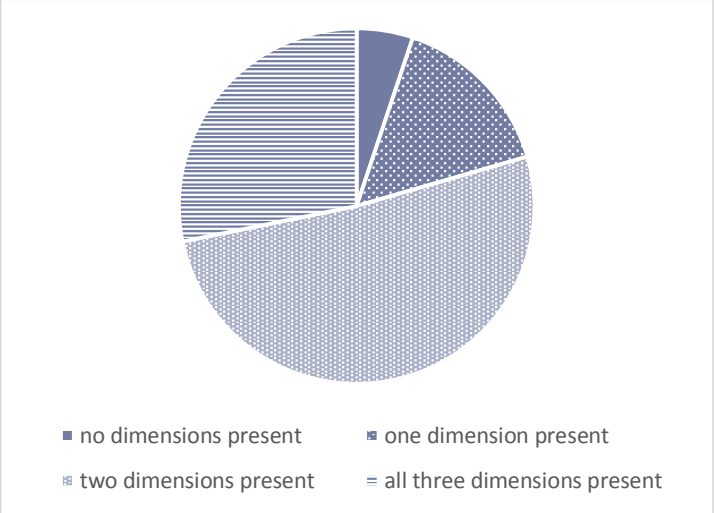


Figure 7: Quality of evidence in the set of 39 evaluations

Limited data triangulation

Nearly 44% of the evaluations based most or almost all their conclusions on data from a single stakeholder group (most frequently, those implementing the intervention). Only 18% of the evaluation teams triangulated the perspectives of active stakeholders in the intervention (implementers and donors) with those of beneficiaries and non-stakeholders.

Selection bias

Some 58% of the evaluations had a strong or very strong potential for selection bias. The potential for selection bias was considered high if, for instance, interviewees were chosen by

implementing organisations. Another example would be selecting interviewees only among stakeholders who strongly benefitted from the evaluated intervention.

Power bias

In our research, power bias referred to the settings and organisation of interviews and group discussions. For instance, where women were interviewed in the presence of men, gender inequality probably made it difficult for them to express themselves freely. In such situations, we rated the potential for power bias as high. Strong or very strong potential for power bias was found in 31% of the evaluations.

Extent to which conclusions were based on original data				
Almost none				2,6%
Some				2,6%
Most				30,8%
Almost all				64,1%
Data triangulation: *Most or almost all conclusions based on data from…*				
Either active intervention stakeholders or beneficiaries				43,6%
Active intervention stakeholders as well as beneficiaries				38,5%
Active intervention stakeholders, beneficiaries as well as non-stakeholders				17,9%
Potential bias in data collection	***absent***	***weak***	***strong***	***very strong***
Selection bias	19,4%	22,6%	35,5%	22,6%
Power bias	62,1%	6,9%	13,8%	17,2%
Aspects of transparent documentation				
Terms of Reference provided				43,6%
At least one data collection tool documented				53,8%
Discussion of limitations of the approach				53,8%

Table 4: Quality of evidence in 39 reports

Gaps in documentation

For almost half of the evaluations, we noted gaps in the documentation of the framework of the evaluation and its tools: Some 56% did not include the evaluation TOR; 46% did not document any data collection tool; 46% did not discuss any difficulties and limitations the evaluators had experienced. That made it difficult for readers to reconstruct the basis on which the evaluation was carried out, and to assess whether the findings were based on accurate data.

4.3.3 Communication

We assessed how evaluation findings were shared. The way in which information was presented varied, with most reports scoring highly. In some 67% of the evaluations, the findings, conclusions and recommendations were stated in an accessible manner. The table below shows individual aspects of layout and content that have facilitated access and understanding of the information.

Distribution of evaluation reports was quite restricted. All reports were shared directly with evaluation stakeholders (via e-mail, hard copy, personal presentation or a combination of these options). But less than 19% were published via several on-line media. This low-cost, yet potent distribution channel could have been used more extensively.

Aspects of presentation			
Executive summary or equivalent is present			89,7%
Of which: Executive summary presents findings, conclusions, recommendations and lessons learned in a way understandable to development practitioners			93,9%
Accessible language			82,1%
Layout: Key terms			38,5%
Layout: Interview quotes			70,6%
Layout: Informative inserts			44,7%
Layout: Subheadings			97,4%
Aspects of distribution	**None**	**Weak**	**Strong**
Direct distribution of report (e-mail, hard copy or personal presentation)	0,0%	38,5%	61,5%
Dissemination of report via list servers, websites and/or social media	3,1%	78,1%	18,8%
Sharing of evaluation findings in other documents and/or workshops	0,0%	36,1%	63,9%

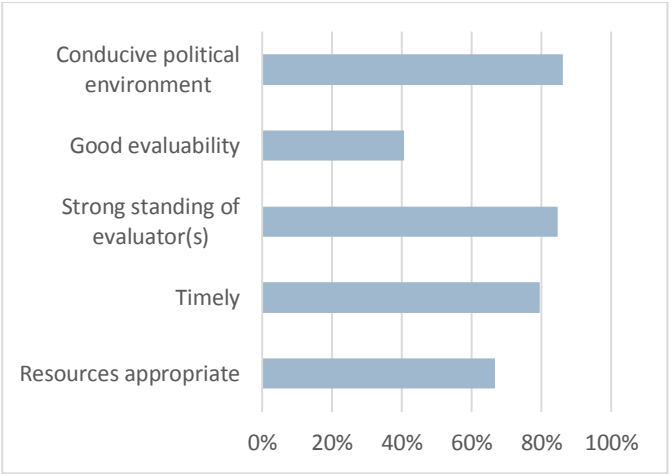
Table 5: Presentation and distribution of 39 reports

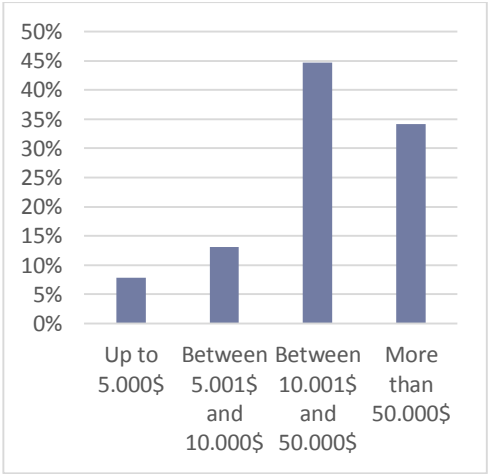
4.4 Evaluation context

We defined evaluation context as a combination of factors related to (i) the evaluated intervention (clarity of design, availability of data), (ii) the evaluation task, (iii) the resources reserved for the evaluation (time, funding and skilled, independent evaluators), and (iv) the situation of the organisations involved in the intervention.

Our survey showed that 31 out of 39 evaluations had occurred in a favourable political environment: the organisations involved had an interest in learning from the evaluation, and enjoyed relative continuity in terms of staff and funding. In 33 cases, the evaluators had high standing in terms of professional skills and independence. Only slightly more than 25% of the evaluators stated that their budget was inadequate in view of the evaluation tasks.

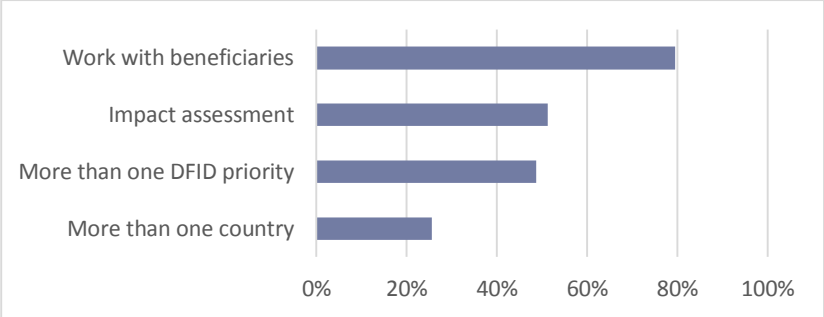
Gaps in evaluability were fairly common, affecting almost 60% of the examined evaluations. Only 14% of the terms of reference (TOR) included an adequate definition of the intended beneficiaries of the intervention; 32% stated specific objectives of the intervention; and 22% had an explicit theory of change. Lack of baseline data or other previous research on the intervention was also a problem in 31% of the evaluations.





Figures 8 and 9: Evaluation context and resources in the set of 39

Some 49% of the evaluation teams faced a **complex evaluation task**, measured on the basis of four criteria related to the intervention and to the evaluation task.



We considered evaluation tasks to be complex if they included an impact assessment. Other factors: number of countries the intervention covered, number of DFID priorities (i.e. themes) and whether the intervention directly engaged with 'ultimate beneficiaries'.

Figure 10: Complexity of task

4.5 Effects of evaluations

We assessed four types of effects caused by evaluations. Three types of effects focused on active stakeholders, i.e. the organisations that implemented or funded the intervention (or other interventions in the field of VAWG):

- **Action effects:** The evaluation helped to change or reinforce the implementation of an intervention. Such effects could occur (i) at the level of the evaluated intervention (e.g. in a mid-term review); (ii) in follow-up work; or (iii) in wider development practice by those who had implemented the intervention, funders and others working in the field.
- **Persuasion effects:** The evaluation convinced others to support the evaluated intervention (for example, donors maintain or increase their funding), or the policies it advocated for.
- **Learning effects:** The evaluation generated insights and influence affecting the **wider** communities in the fields of development, women's rights and evaluation, beyond and independently from the intervention under evaluation.

Finally, at the level of the women, men, girls and boys who were supposed to benefit from the intervention, an **empowering** effect occurred if, as a result of the evaluation, those **beneficiaries** were consulted more frequently and their voices were heard more forcefully.

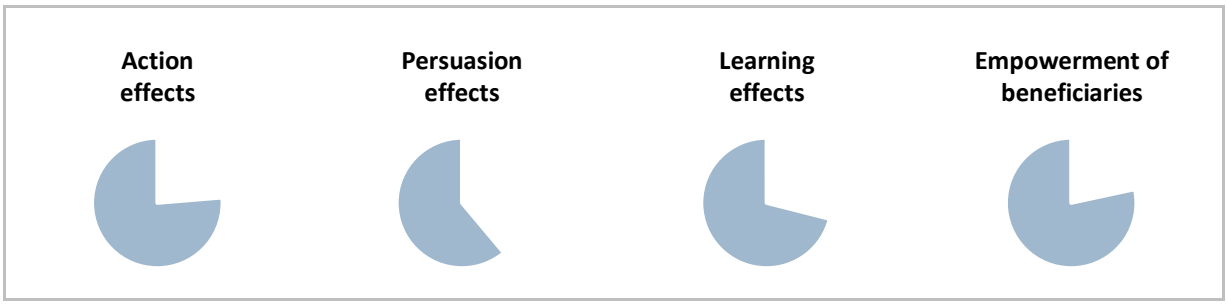


Figure 11: Evaluation effects in the set of 39

There were **strong correlations between evaluation effects**: If an evaluation produced action effects, it often came with persuasion, learning and empowerment as well. The weakest association was between persuasion and wider learning effects.

Effect type	Action	Persuasion	Learning	Empowerment
Action		74%	76%	93%
Persuasion	91%		68%	86%
Learning	82%	58%		77%
Empowerment	86%	64%	71%	

Table 6: Correlations between evaluation effects

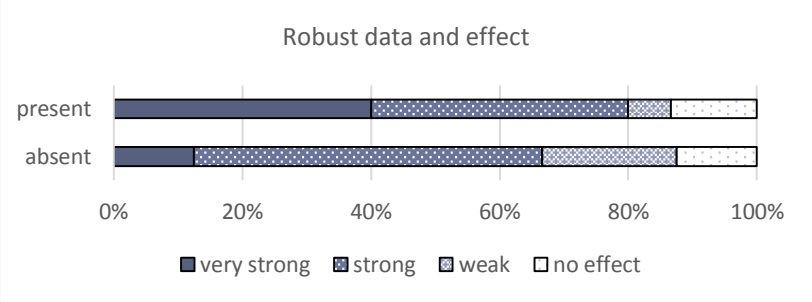
Read the table above from left to right only: For instance, if an evaluation had a strong action effect, in 74% of the cases it had also a strong persuasion effect.

Factors linked to evaluation effectiveness

QCA is the analytical method we have used to identify configurations of factors causing effective evaluations. The results of our QCA are presented in chapter 5 below. Meanwhile, quantitative analysis of the data we gathered revealed interesting associations between evaluation effects and individual characteristics of evaluations in the set of 39 evaluations.

Robust data

Evaluations that based their conclusions on original data, triangulated data sources and avoided selection and power bias in their research design produced strong effects more frequently than evaluations that lacked such robust data.



Evaluations with robust data (top bar) yielded very strong effects to a much larger proportion (40%) than those without robust data (bottom bar)

Figure 12: Robust data and evaluation effects

Sensitivity to evaluation-related risks

With regard to the 39 evaluations in our set, evaluators whose understanding of evaluation-related risks was similar to evaluation stakeholders' perception of those risks (especially risks for direct beneficiaries of the intervention), produced more evaluations with strong effects than less risk-sensitive evaluators.

Dark bars (to the left) represent the percentage of evaluations with strong effects.

The lightest bars (bottom bar and right ends of the two top bars) refer to evaluations that have reportedly generated no effects.

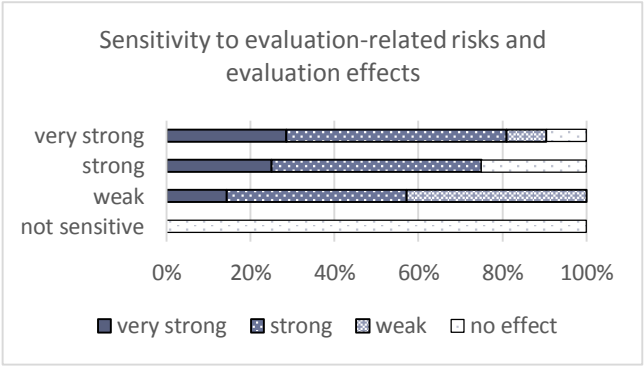


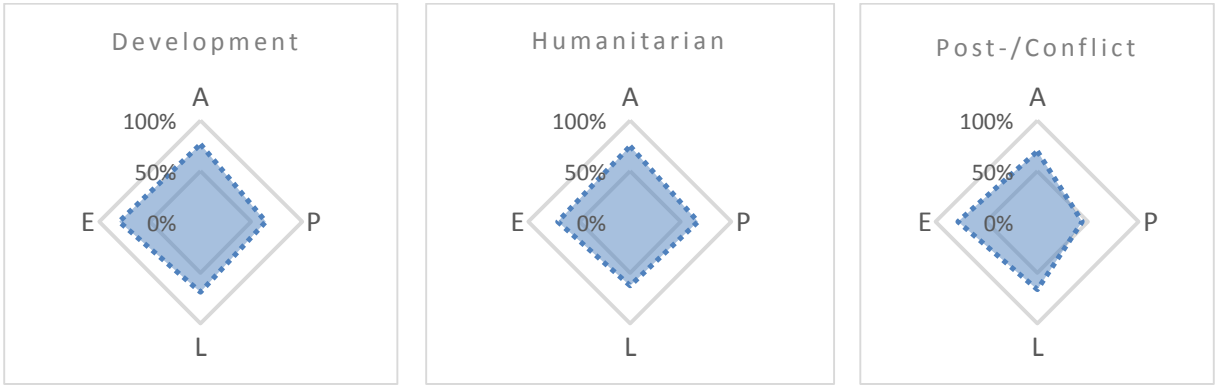
Fig. 13: Risk sensitivity and evaluation effects

Thematic focus of the intervention

We analysed the effects of the 39 evaluations in relation to the DFID thematic priorities on VAWG. Evaluations of interventions in the fields of "empowering of women and girls" and "changing social norms" produced strong persuasion effects more often than evaluations of interventions on "building political will" or "providing services". At the same time, evaluations of interventions on "empowering women and girls" produced strong learning effects (i.e. learning beyond the intervention stakeholders) less often than evaluations of interventions that (also) pursued objectives related to other DFID priorities.

Context of the intervention

Action, learning and empowerment effects occurred most frequently in development contexts. They were slightly weaker in humanitarian and (post-) conflict settings. Evaluations of interventions in (post-) conflict contexts produced strong persuasion effects less frequently than evaluations in other contexts. This may be linked to the risks associated with work in such environments which probably make it harder to convince other actors to support the intervention or its goals.



Acronyms: A= action effect; P= persuasion effect; L= learning effect; E= empowerment of beneficiaries

Figure 14: Intervention contexts and evaluation effects

Evaluation budget

Budget alone did not seem to determine the strength of evaluation effects to any significant degree. Evaluations with a very large budget (>US$ 50.000) produced effects more often than evaluations with a very small budget (<US$ 5.000). But among evaluations with a small budget (US$ 5.001 -10.000$), the share of evaluations with very strong effects was larger than among evaluations with a very large budget.

5. Conditions for evaluation effectiveness

This chapter presents our findings from Qualitative Comparative Analysis: the *paths* or configurations of factors that have made evaluations in the field of VAWG effective. Section 5.1 sketches our definitions of the dimensions of evaluation practice. Section 5.2 presents the *paths* to effective evaluations that we have identified. Finally, section 5.3 illustrates these *paths* with five brief case studies obtained through Process Tracing.

5.1 Conditions for evaluation effectiveness

We have defined seven broad *conditions* for effective evaluation: (i) favourable context, (ii) strongly qualitative and (iii) strongly quantitative approach, (iv) participatory design, (v) sensitivity to the GBV context, (vi) compelling evidence and (vii) good communication.

Information on the distribution of dimensions for each of the conditions in the set of reviewed evaluations is available in annex 2.

Favourable context

This condition brings together three dimensions that relate to the setting in which evaluations are conducted. The dimensions are made up of several contextual factors.

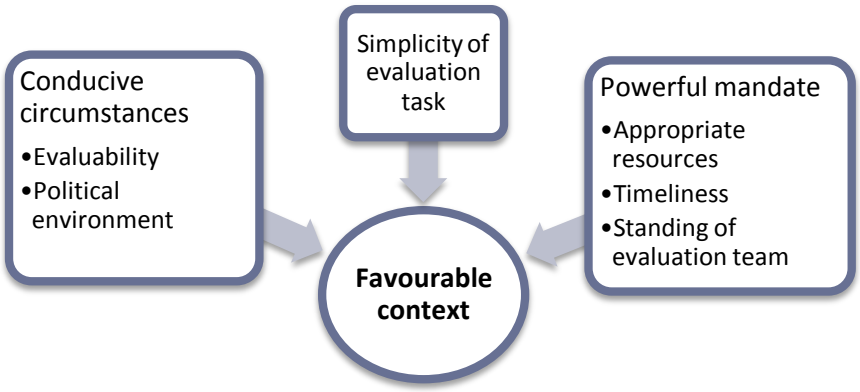


Figure 15: Definition of favourable context

Conducive circumstances

This dimension includes (i) evaluability – i.e. clear intervention design, availability of data – and (ii) a favourable political environment – i.e. evaluation stakeholders who are willing to learn, stability regarding staff and funding, and a reasonably stable external situation.

Simplicity of evaluation task

No evaluation is simple, but evaluations that do not assess impact are simpler than others. The number of countries and/or themes covered, and whether the evaluation works directly with beneficiaries or at a meta-level (on funding policies) also determine the relative simplicity or complexity of the task.

Powerful mandate

This term designates the authority evaluators draw from (i) their professional standing, (ii) the resources available for the evaluation and (iii) the time when it takes place.

Favourable context factors can partially compensate for unfavourable context factors (i.e. factors that are not conducive to effective evaluation). For instance, where the evaluation task is highly complex, the evaluation can still be effective, provided the evaluators' mandate is powerful, the intervention design is clear, and the political environment is stable. We have therefore taken the mean of the three dimensions as the value for "favourable context".

Approach

We distinguish between qualitative and quantitative approaches (see 4.2 above). Individual conclusions of the evaluation report can be backed by quantitative and qualitative data; hence some evaluations are both 'strongly quantitative' and 'strongly qualitative'.

Compelling evidence

We have separated the choice of data collection methods ("approach", above) from the evaluators' compliance with standards of scientific research ("compelling evidence").

Compelling evidence rests on two pillars:

(i) Robust data

Data was gathered according to scientific standards accepted across the qualitative/ quantitative divide (see section 4.3 above for detail).

(ii) Transparent documentation

Documentation of the research process was transparent if data collection and analysis were detailed in a way that allowed others to replicate the research process.

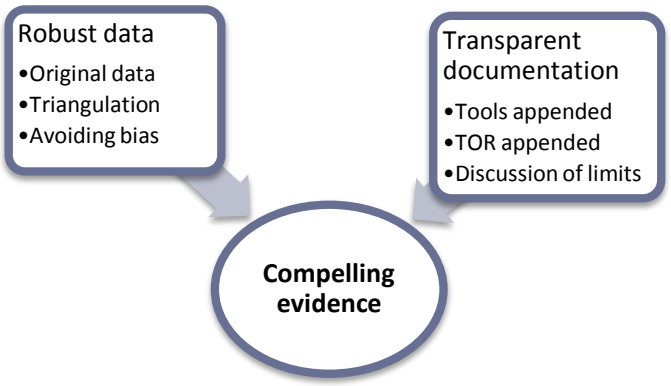


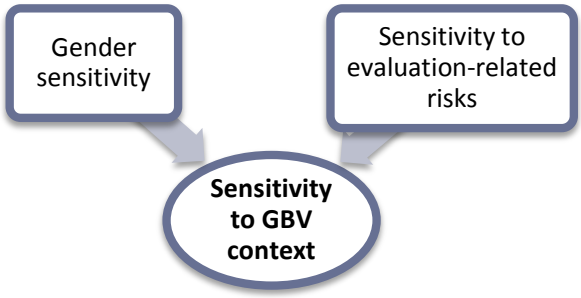
Figure 16: Definition of compelling evidence

We assume that if either of the dimensions outlined above is seriously flawed, the evidence will not lead to evaluation effectiveness. Therefore we have used the minimum of the two dimensions as the overall value for 'compelling evidence'.

Sensitivity to GBV context

This condition is about the evaluators' practical understanding of two main aspects of research in the field of violence against women and girls: (i) gender sensitivity and (i) sensitivity to the risks linked to the evaluation for informants, researchers and others (for example, security risks and risk of re-traumatisation).

Figure 17: Definition of sensitivity to GBV context



We considered both dimensions described above to be necessary for an evaluation process sensitive to the GBV context. For instance, an evaluation team that was highly knowledgeable about gender would only be categorised as "sensitive to the GBV context" if they were also sensitive to the risks an evaluation could present for the rights and well-being of its participants.

Participatory design

In our definition, participatory design means that intervention stakeholders (those implementing the intervention, donors and intended beneficiaries) have played an active role in evaluation design and data analysis.

We measured this condition by asking evaluators and representatives of the organisation implementing the evaluated intervention whether workshops were held near the beginning and the end of the evaluation, to discuss evaluation design and the findings respectively.

Good communication

Good communication was defined as a result of (i) wide **dissemination** of the evaluation findings, conclusions and recommendations (distribution) and (ii) accessible **presentation** in the report. For more details on these dimensions, see section 4.3 above.

Figure 18: Definition of good communication

We consider both clear presentation and wide distribution to be necessary elements of good communication. Therefore we took the minimum of the two dimensions as the value for the aggregate.

5.2 QCA findings: configurations for effective evaluation

5.2.1 Evaluation effects as *outcomes*

The effects of evaluations are the *outcomes* of our QCA. Section 4.5 above shows that most evaluations in the set of 39 have generated at least one type of strong effect.

Implications for our analysis. Evaluation purposes vary and the likely effects of an evaluation depend to some extent on the evaluation purpose. Therefore, we treated the three types of **effects on evaluation stakeholders** (action, persuasion, and learning) as substitutes. When determining the value to be used for the QCA, we used the highest value achieved among the three types of effects, i.e. the maximum.

Regarding the **effects on the intended beneficiaries** of the intervention, we believe that **an evaluation cannot be deemed effective if beneficiaries were harmed in the process**. Therefore, when assessing the overall effects of an evaluation, we used the minimum of the effects on beneficiaries and other stakeholders – i.e. the evaluation had to display strong effects on both sides to be considered effective.

5.2.2 Paths to effective evaluation

Qualitative Comparative Analysis yielded a complex solution of 11 *paths* composed of five or more individual conditions, based only on configurations observed among the evaluations with strong effects. Using simplifying assumptions, we obtained eight *paths* leading to evaluation effectiveness. Together, the 8 paths explain nearly 90% of the evaluations' effect strength.

For a decision-tree presentation of the paths, please refer to the executive summary and chapter 6 below.

The table on the following page displays the conditions that shape the eight individual paths. The column to the right shows the percentage of cases with strong effects that the respective configuration represents[11].

[11] Percentages add up to more than 100% because some cases are covered by more than one path.

Path	Approach		Context	Evaluation quality				Cases covered (%)
	Strongly qual.	Strongly quant.	Favour. context	Particip. design	Compell. evidence	Good comm.	Sensitive to GBV	
1								53,6%
2								21,4%
3								17,9%
4								10,7%
5								10,7%
6								28,6%
7								35,7%
8								7,1%

Table 7: Sufficiency paths of the intermediate solution

Colour coding: Grey colouring means that the *condition* must be present as part of the respective *path*. Red means that the respective *condition* must be absent from the *path*. If a *condition* is neither red nor grey, it does not matter whether it is present or absent for the path to lead to effective evaluation.

Paths to effectiveness for evaluations with strongly qualitative designs: Paths 1, 3, 6, 7 and 8 above produced effective evaluations using designs that were almost exclusively quantitative. Path 1 covered some 54% of cases with strong positive effects, which resulted from a strongly qualitative and participatory design executed by highly sensitive evaluators.

Three of these *paths* (6, 7, 8) required a favourable context. For these *paths,* the addition of participatory design, compelling evidence or high sensitivity to the GBV context was sufficient to produce strong effects. **In non-favourable contexts, qualitative evaluation required participatory design as a *necessary condition* for effectiveness** (*paths* 1 & 3).

Path for strongly quantitative designs: In evaluations with conclusions almost exclusively based on quantitative data collection, *paths* 2 and 5 generated effectiveness. Such evaluations required a participatory design, regardless of the evaluation context. In a non-favourable context, high sensitivity to the GBV context was also a *necessary condition.*

Paths for strongly mixed designs: For evaluations that based most or almost all conclusions both on qualitative and quantitative data, *paths* 1, 2 and 4 produced effectiveness. In all three *paths*, sensitivity to the GBV context was a *necessary condition*; the evaluation context was unimportant.

Insights from QCA without "approach" as a condition

We conducted an additional QCA that included evaluation context and the four dimensions of evaluation quality only (i.e. without taking into account the choice of qualitative or/and quantitative approaches).

The importance of **participation** and **sensitivity to the GBV context** was confirmed by the *paths* in the *intermediate solution.* The presence of those two factors alone covered 61% of the evaluations with strong effects. The two *paths* with the next highest coverage of cases with strong effects (50% and 43% respectively) showed that in favourable evaluation contexts, either of those two conditions was *sufficient* for effectiveness.

5.3 Case studies: four paths to effective evaluation

QCA shows the configurations of *conditions* that lead to *outcomes*, but does not reveal how they bring about those outcomes. We used Process Tracing to learn about the interplay of conditions in five cases that stood for four different methodological choices.

5.3.1 Quantitative methods, participatory design and favourable context: Kim *et al* (2009)

A particularly effective evaluation representing *path* 5 is *"The Refentse Model for Post-Rape Care: Strengthening Sexual Assault Care and HIV Post-Exposure Prophylaxis in a District Hospital in Rural South Africa"* by Julia C. Kim, Ian Askew, Lufuno Muvhango, Ntabozuko Dwane, Tanya Abramsky, Stephen Jan, Ennica Ntlemo, Jane Chege and Charlotte Watts. It was designed in parallel with the actual intervention. Its main features are outlined in the annex (short descriptions of evaluations); the full report can be downloaded from the Population Council website.[12]

The configuration

The evaluation represents *path 5* , which is made up of four conditions, found to be jointly sufficient to bring about positive evaluation effects:

(i) favourable context for the evaluation, (ii) participation, (iii) strong role played by quantitative methods, and (iv) a weak role for qualitative methods[13] when producing the evidence for the evaluation conclusions (as presented in the report).

We found the same configuration in two other evaluations of our QCA set: CARE 2009 and Rujumba 2012.

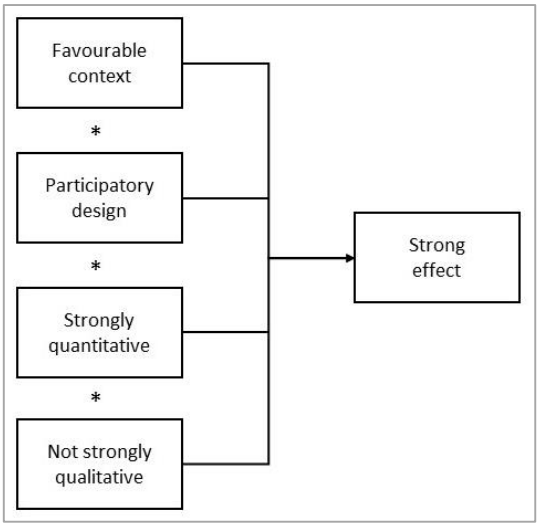


Figure 19: Sufficiency path for Kim (2009)

Key evaluation findings

The evaluation demonstrated that rural hospitals could provide effective care for rape survivors, in particular post-exposure prophylaxis (PEP), with the staff and infrastructure available in resource-poor settings. Action research supported the development of a practical, nurse-driven model for post-rape care.

[12] http://www.popcouncil.org/uploads/pdfs/frontiers/FR_FinalReports/SouthAfrica_RADAR.pdf (23/4/2014)

[13] The conclusions presented in the evaluation report are based on quantitative data. However, most of our interlocutors spoke of a mixed methods approach, as the continuous presence of researchers provided many occasions for observation and discussion. Rather than describing this as a qualitative methodology, we would see the approach as presenting rich opportunities for participation.

Pathways to success in the Kim (2009) case

Note: Terms that are underlined refer to QCA conditions and effects identified in the Review. Double underlining refers to the top-level conditions used in QCA, and single underlining refers to their components.

Favourable context: The intervention was designed as action research, i.e. with the intention to test a model of post-rape care. It came with an explicit theory of change and baseline data – the two prerequisites for evaluability. The evaluation task could be qualified as simple in that it focused on a single hospital and was closely connected to daily hospital work. This favoured the development of a robust design which readers of the evaluation could understand and apply the findings in their practice. *"It was quite a functional straightforward design that […] when it was translated into results, people could get their heads around pretty quickly."* (Kim 2009 #3)

Appropriate resources were available, allowing members of the evaluation team to work on-site throughout the intervention. That strengthened the participatory aspect of the research. The team displayed high levels of professional standing and independence, bringing together international researchers from reputable institutions with experienced local researchers. This set-up secured high quality data gathering and analysis, as well as dissemination of findings through the researchers' professional networks and in a peer-reviewed journal.

Local rootedness as a special case for participatory design

Local and international evaluators were at the intervention site throughout the intervention process. Frequent consultation between researchers and practitioners helped to fine-tune data collection instruments and data analysis, and to continuously feed findings into the intervention. *"For me as a researcher the experience of living in that community and being based there for some time while the research happened was really important on the design side of things. Just making sure that the research questions and the design and the way of going about it was appropriate to the environment. […] You know, you have your data and you come to conclusions based on that. But there is a lot of more subtle information that comes from being there, knowing the context and understanding the difficulties. […] It helps me to interpret the data and maybe to present it in a way that is relevant and maybe a bit more authentic. […]*

At periodic points we would review how all the data collection was going […]. I think we had to, as a team, to quite commit to the issue and to following up with the hospital and the group that were involved in the study, to make sure that the quality [of data collection in hospital records] was good." (Kim 2009 #3)

The participation of South African researchers was also seen as essential: *"Because this evaluation was undertaken by the resident researchers who were […] on-site 24/7. They weren't coming in from Johannesburg, or Europe or America; they were there, so they had a good understanding of what was happening; they helped to adjust things as it moved along."* (Kim 2009 #1)

The evaluation was timely, as post-exposure prophylaxis (PEP) had become available and there was strong interest in its applicability both in South Africa (where a national policy on PEP distribution had been introduced) and among international donors.

Furthermore, the political environment of the evaluation was stable in that (i) main actors in the intervention were present throughout its implementation, and (ii) no external disruption was reported. The South African NGO partner worked in an area with high HIV prevalence; developing effective PEP for rural settings was a central issue for them. The study had been commissioned by the Population Council, described by our interviewees as *"not so much the donor or the funder than a kind of technical partner"* (Kim 2009 #3).

Participatory design: The evaluation was highly participatory, involving a broad spectrum of actors ranging from hospital staff to provincial and national representatives of Health authorities and other sectors.

Benefits of multi-sector cooperation

In Kim *et al.* (2009) the evaluation contributed to bringing together key actors, fostering a dialogue on more effective collaboration.

"Because it was a multi-sector intervention involving police, involving social workers […] - they were not groups that met together necessarily before. […] So we would have meetings as part of this committee where the different sectors came together. And I think that was important because I think often-times people did not really know what the other players were doing or had assumptions about it. […] But when people started to meet more regularly, I think it built that sense of understanding. So that kind of participation was important for the intervention success in itself. You know, you have guidelines for what police should do and you have guidelines for what social workers should do. But then to actually sit together periodically and talk and say 'This is why. This is what is difficult for us. Can you guys try and do this? Why can we not keep the rape kit in the examining room instead of at the police station?" (Kim 2009 #3)

Quantitative method: Last but certainly not least, the strong use of quantitative data in the presentation of findings played a major role in generating positive evaluation outcomes. As one interlocutor put it: "*We needed [data on what] quantitatively changed the service - we had more women coming, we had more women getting effective care, we were receiving quantifiably a higher level of qualitative care. So, all of these three outcomes had to be measured quantitatively to convince decision makers that this was the way to go. […]"* (Kim 2009 #1).

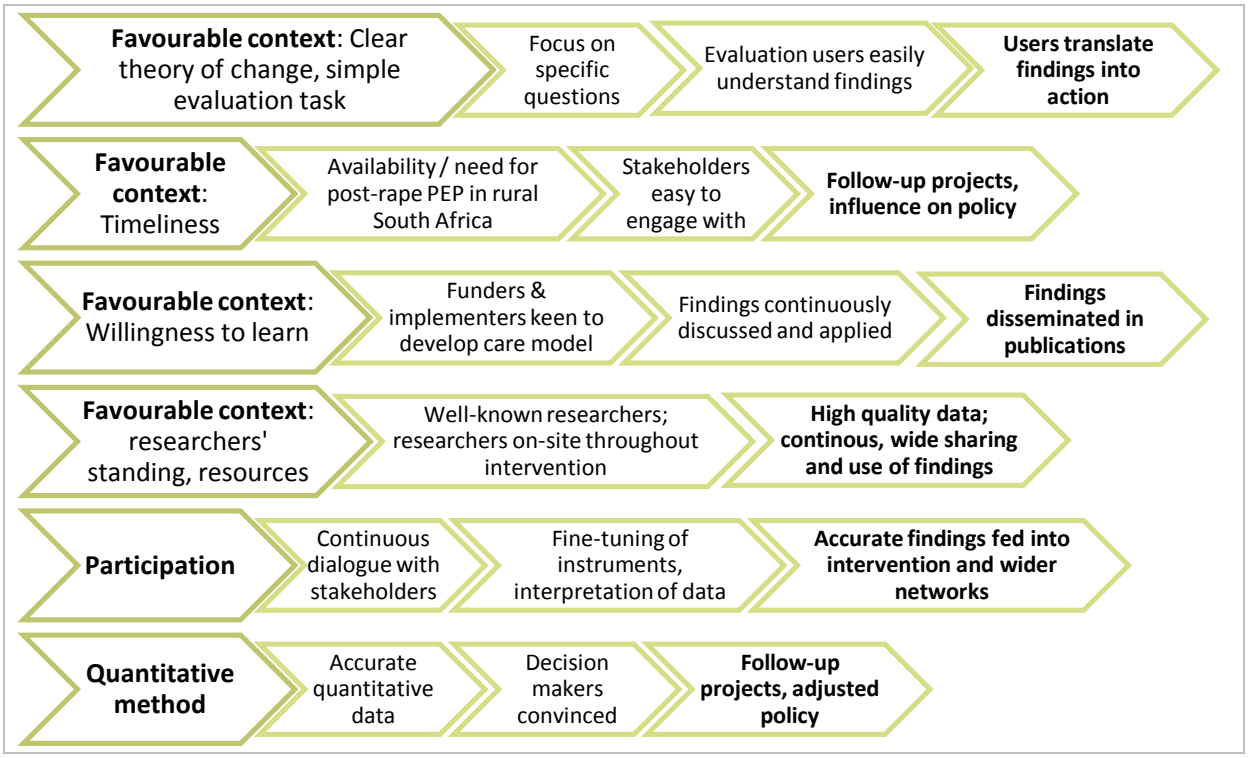


Figure 20: Cause-to-effect chains for Kim (2009)

Evaluation effects

The evaluation brought about strong action effects: it led to a second and third phase of programme implementation; its findings were used in capacity building across the South Africa region. The study was published and disseminated widely, which increased potential wider learning effects. To a lesser degree, persuasion effects have been observed in that the study contributed to expanding stakeholders' networks linked to the intervention.

Another typical case for this path: Rujumba (2012)

The *"Midterm Review of the We Can Campaign (WCC) to End all Violence against Women"* in Uganda, commissioned by Oxfam in Uganda and carried out by the Agency for Capacity Building (ACB), displayed the same *path* (5) as Kim (2009). ACP ran a household survey to (i) assess campaign effectiveness and (ii) identify baseline information for future implementation phases. Furthermore, interviews and discussions were held with VAWG survivors and campaign alliance members. (For more details, see annex.)

The mid-term review found that the campaign had reached a third of its target population. It provided recommendations for the handover of campaign management from Oxfam to the campaign alliance.

Pathways to success in Rujumba (2012)

Favourable context: The evaluation happened shortly before campaign management was transferred from Oxfam to the campaign alliance, i.e. it was timely and the stakeholders' willingness to learn was strong. The task was relatively simple (a single intervention in a single country, focusing chiefly on behaviour change). The evaluators enjoyed high standing, being independent and from Uganda, i.e. highly knowledgeable about Ugandan society.

Participation: Workshops were held near the beginning and the end of the review. Furthermore, the evaluator interacted with a large number of campaign alliance members when interpreting data and developing recommendations. As in Kim (2009), such close contact with key stakeholders fostered their ownership of the evaluation and led to highly relevant findings and recommendations.

Method: The predominantly quantitative design produced an accurate snapshot of knowledge, attitudes, practice and beliefs related to VAWG at grassroots level, and created a sense of urgency for the campaign. *"[Campaign alliance members] are now feeling like obliged to do much more than what they had thought they should be doing, because whatever the evaluation came up with, those words were coming from the people that we were trying to support."* (Rujumba 2012 #2)

Evaluation effects

The evaluation generated strong action effects: it reportedly reinforced the campaign alliance and informed its subsequent work plans. In particular, the creation of formal "change maker circles" (campaign multipliers), constructive engagement with men and boys, and work in schools were strengthened. Funding to the campaign was maintained (persuasion effects).

5.3.2 Qualitative methods, participation and sensitivity: Robinson (2011)

An example for *path* 1, followed by 54% of the effective evaluations in our QCA set, is the evaluation *"Putting the Jigsaw together – CARE International Sri Lanka's Violence against Women Intervention in Batticaloa: 2003-2011"* by Victor C. Robinson in Sri Lanka. It was based on key informant interviews, group discussions, direct observation and review of programme documents.

The evaluation occurred near the end of an 8-year sequence of interventions to end gender-based violence. A respondent described it as *"a history rather than an evaluation*", as one-third of the evaluation report narrates the successive CARE interventions against gender-based violence in Sri Lanka. A section with the title "reflection" reconstructs the theory of change, discusses effectiveness and impact, and presents suggestions for subsequent work.

Key evaluation findings

The evaluation found CARE's interventions to end VAWG effective. It recommended continued CARE programming on gender-based violence, which in Batticaloa had evolved into a model *"firmly rooted in Sri Lankan Culture and experience"* (Robinson 2011: 32). The evaluator stressed the importance of a theory of change based on social analysis, and recommended the establishment of systems to measure social change and to foster learning.

The configuration

Path 1 included three conditions, which were jointly sufficient to bring about positive evaluation effects: (i) participation, (ii) sensitivity, and (iii) qualitative approach.

Other evaluations in our QCA set that followed *path* 1 include: Carty 2009, Chibuta 2011, Creighton 2011, Diop 2008, FASI 2011, Germann 2010, Harvey 2012, Ingdal 2008, Naik 2010, Naik 2012, Odhiambo 2011, Pittman 2010, Robinson 2012 and Shaheed 2011.

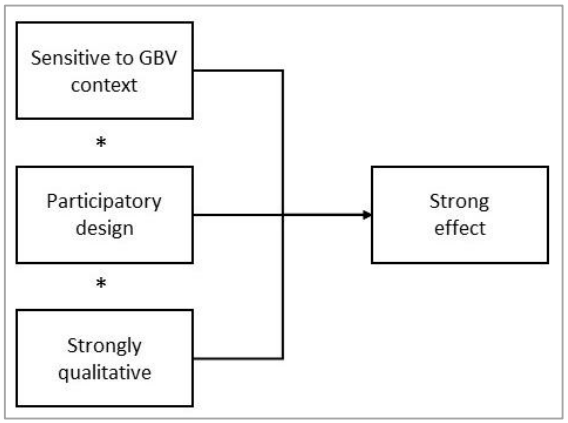


Figure 21: Sufficiency path for Robinson (2011)

Pathways to success in the Robinson (2011) case

Participation: The evaluation was highly participatory, engaging a wide range of staff and programme partners in joint reflection. It included workshops at the beginning and near the conclusion of the field work. *"If you are going to learn in the evaluation, then you get everybody involved in learning, and so even if I am doing focus group discussions with beneficiaries, with target groups, I am asking them, 'what can we learn from this project, you tell me what we can learn'"* (Robinson 2011 #1)

Arguably, the qualitative method fostered such participation, creating venues for feedback by implementers at various points. *"What I would do is, to sometimes formally, sometimes informally, gather these people from the field and we'd talk about what happened, what lessons we pulled from that."* (Robinson 2011 #1)

The qualitative, explorative approach matched the purpose of the evaluation, which was intended to generate a fuller understanding as to how CARE's VAWG programme worked. The interviews and group discussions generated real life stories, which our interlocutors found to *"speak to people more easily"*. (Robinson 2011 #1) The effort to get across the findings in an emotionally engaging way was reflected in the presentation of the report, structured like a historical narrative. Reconstructing the programme theory of change, the evaluation provided the implementing organisation with a useful instrument for subsequent planning and monitoring. *"Even today, we'd say, 'go back to what [the evaluator] wrote'; because he suggested that our work has three pieces to it […]: normative change, structural change and policy change. These three words, we keep using them […] We did not have that before [the evaluation] was done."* (Robinson 2011 #2)

Gender sensitivity and ethics contributed to generating the trust needed to obtain rich and authentic responses. In addition to the conditions in this solution *path*, the evaluator's familiarity with the national context and with CARE was quoted as a major source of trust and a factor for evaluation success.

Evaluation effects

The evaluation reportedly brought about strong action effects: it convinced CARE Sri Lanka to continue VAWG programming. It was shared across CARE International, in full and as a summary in the CARE newsletter. Furthermore, the evaluation generated persuasion effects in that a CARE International affiliate provided additional funding for related work on VAWG, and convinced donors to back long-term, flexible approaches in VAWG programming. No wider learning effects beyond CARE were reported to us, which was probably linked to the fact that the evaluation remained unpublished.

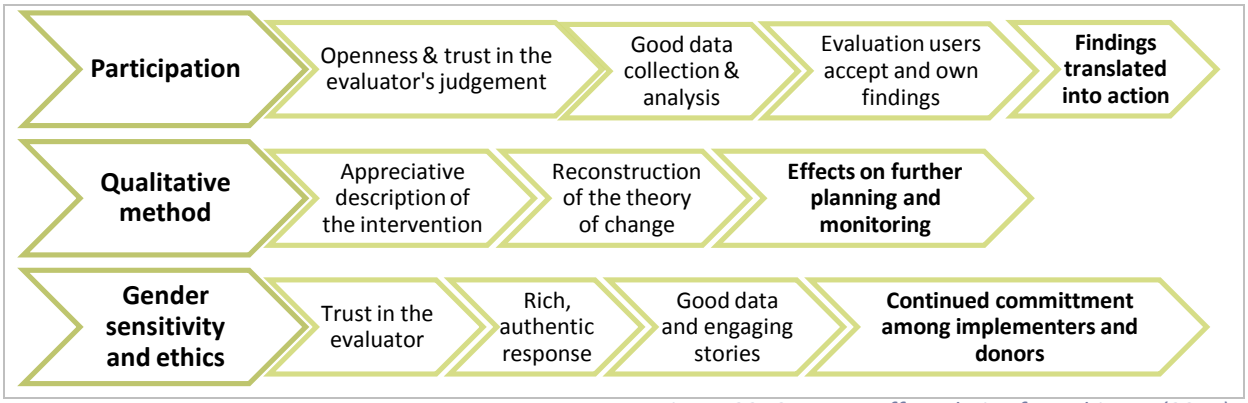


Figure 22: Cause-to-effect chains for Robinson (2011)

5.3.3 Qualitative method, sensitivity and favourable context: Moen et al. (2012)

The configuration

The *Comparative Evaluation of FOKUS FGM projects in East Africa* by Hanne Lotte Moen, Aggripina Mosha and Hirut Teferi stood for *path* 7, composed of four conditions which were jointly sufficient for positive evaluation effects: (i) favourable context, (ii) sensitivity, (iii) strong role for qualitative methods and (iv) minor role for quantitative methods.

We found the same configuration in 9 other evaluations from the QCA set: Chibuta 2011, Diop 2008, Fawzi 2011, Ingdal 2008, Naik 2010, Robinson 2011, Robinson 2012, Sotirovic 2012, Townsend 2010.

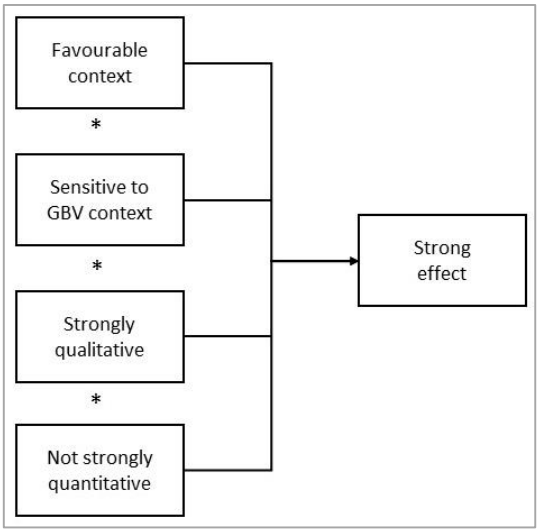


Figure 23: Sufficiency path for Moen (2012)

Key evaluation findings

The evaluation found that all FOKUS projects had generated public debate on FGM; and that they had probably contributed to a reduction in FGM in several areas. Local ownership, long-term work and targeting of key actors were identified as key factors for success. The evaluation report included recommendations on the planned regional programme.

Pathways to success in the Moen (2012) case

Favourable context: The evaluation was timely: it came at a moment when FOKUS needed advice on the planned integration of FGM initiatives into one programme. The programme stakeholders reported strong willingness to learn on the part of key evaluation stakeholders. *"FOKUS […] decided that from then on we were going to try to do more thematic evaluation instead of individual project evaluations. So there was this kind of enthusiasm within the organisation to always look at a group of projects that worked on the same topic. […] There was a lot of support internally among our staff; there was support and cooperation not only from our implementing partners but also their Norwegian counterparts."* (Moen 2012 #1).

No major staff changes or disruptions in the wider context were reported; i.e. the political environment of the evaluation was stable. The evaluation team enjoyed high standing in that it included evaluators from Europe, Ethiopia and Tanzania familiar with FGM-related interventions. The lead evaluator was well acquainted with FOKUS, which reportedly made it easy for her to understand the evaluation task and the stakeholders' roles in the intervention. The resources for the evaluation, both in terms of funding and the time contributed by programme staff, were broadly described as adequate. The task was relatively simple: although it covered three countries, a single type of intervention was evaluated; no impact assessment was included.

Gender sensitivity and ethics: The evaluators were sensitive to evaluation-related risks, taking precautions to protect informants.

Qualitative method: Due to the lack of baseline data, qualitative data collection was reported to be the only option for this evaluation. Interviews and group discussions with community members provided rich data that helped appreciating the different methods used in tackling FGM in East Africa. The data was used by a range of stakeholders – FOKUS, intermediary (Norwegian) organisations and implementing local partners – to adjust their programmes. Furthermore, the evaluation provided practical advice regarding themes, implementation and monitoring mechanisms for the planned regional programme.

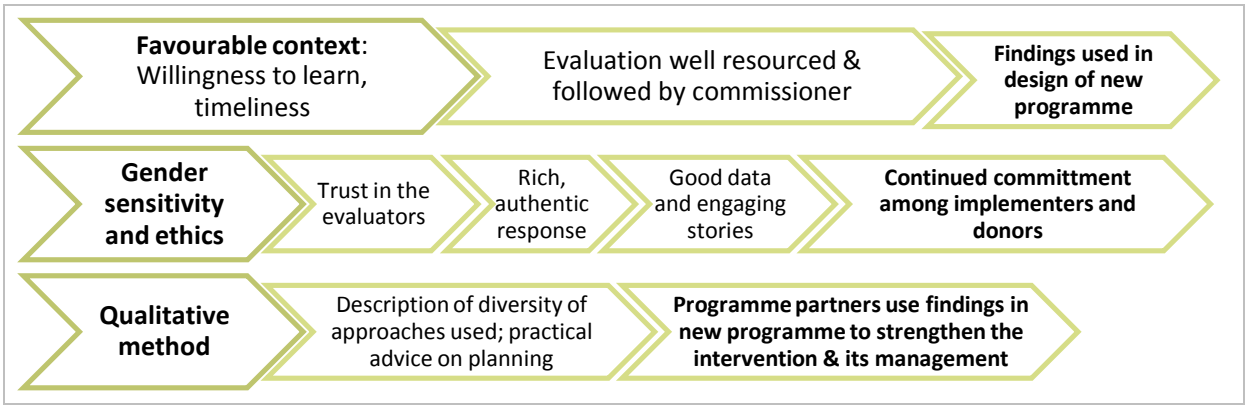


Figure 24: Cause-to-effect chains for Moen (2012)

Evaluation effects

The evaluation reportedly caused strong action effects, prompting FOKUS to establish the regional programme based on the evaluators' recommendations. That included the reduction of countries covered from three to two, and the introduction of a system to monitor results. Persuasion effects were generated in that FOKUS felt encouraged to substantially increase

funding to its Tanzanian partners.[14] Wider learning effects reportedly occurred in the Norwegian development community, where the evaluation report was disseminated.

5.3.4 Mixed approach, compelling evidence and sensitivity: Mwangi (2012)

The configuration

The *Gender Based Violence Program Evaluation* under the CARE Refugee Assistance Programme in Dadaab (Kenya) by Gladys Kabura Mwangi followed *path 4.* The path was composed of four conditions, which were jointly sufficient for positive evaluation effects: (i) compelling evidence, (ii) sensitivity, and (iv) a strongly qualitative and (iv) strongly quantitative approach.

Two other evaluations in the QCA set followed the same path: Harvey 2012 and Marrar 2010.

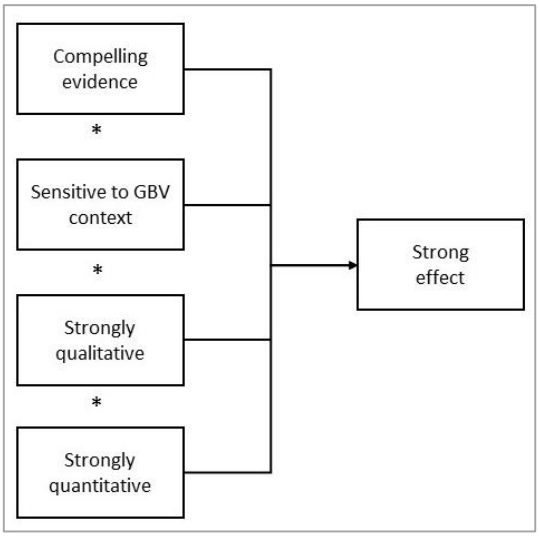


Figure 25: Sufficiency path for Mwangi (2012)

The purpose of the evaluation was to examine CARE's interventions on GBV (2001-2011) in terms of their effectiveness and impact among the refugees in Dadaab, and to provide thematic recommendations for future programming. Its conclusions rested on quantitative and qualitative data collected during the evaluation, including a survey with 400 refugees, focus group discussions, key informant interviews, field observation and text analysis.

Key evaluation findings

The evaluation concluded that some positive attitude and behaviour change regarding GBV and FGM had occurred over ten years, even though law enforcement (for instance, against FGM) remained deficient and medical services were overstretched. The report listed some 30 recommendations on future programme strategy and management, including *inter alia* ideas for advocacy with police and other agencies working in Dadaab. It recommended set up integrated services for GBV survivors, to increase medical and psychosocial services, to involve men in GBV prevention and to strengthen coordination with other actors.

Pathways to success in the Mwangi (2012) case[15]

Mixed approach: The combination of qualitative and quantitative data collection yielded robust figures on the prevalence of GBV (in particular FGM) in the Dadaab camp, as well as examples to illustrate the figures and of the challenges encountered by the programme.

Compelling evidence: According to the evaluation report, the survey was performed to appropriate quality standards, including data triangulation and bias control. The method was transparently presented, as the report included the data collection tools in annex (survey questionnaire, focus group discussion guide, interview guide, sampling frame, enumerator schedule). That made the findings highly credible. Findings were presented at

[14] However, it was perceived that more robust data obtained in a quantitative impact assessment could have persuaded donors to FOKUS, such as the Norwegian government, to increase funding to FOKUS.

[15] Due to difficulties in contacting evaluation stakeholders, only one full interview and a brief e-mail exchange could be organised, yielding relatively little data on the way in which the *solution path* played out.

inter-agency coordination meetings where they reportedly created a sense of urgency, encouraging other agencies to cooperate with CARE on FGM in Dadaab.

Ethics and gender sensitivity: Ethical guidelines were observed so as to prevent any harm participants could potentially experience as a result of the evaluation. All respondents were reportedly informed that participation was voluntary; informed consent forms were used. *"The evaluator had only access to information that was non-identifying information. So even the evaluator would not be able to know who responded."* (Mwangi 2012 #2) The evaluator held debriefing sessions with women or groups who reported to have experienced violence or to know someone who had experienced violence.

Evaluation effects

The evaluation caused action effects, providing findings that CARE used to design follow-up programmes. For instance, recommendations to strengthen work with the police, and with men and boys, were taken up. Persuasion effects occurred in that the findings convinced other agencies in Dadaab to pay greater attention to FGM-related work; one international agency reportedly announced it would cooperate with CARE on the topic.

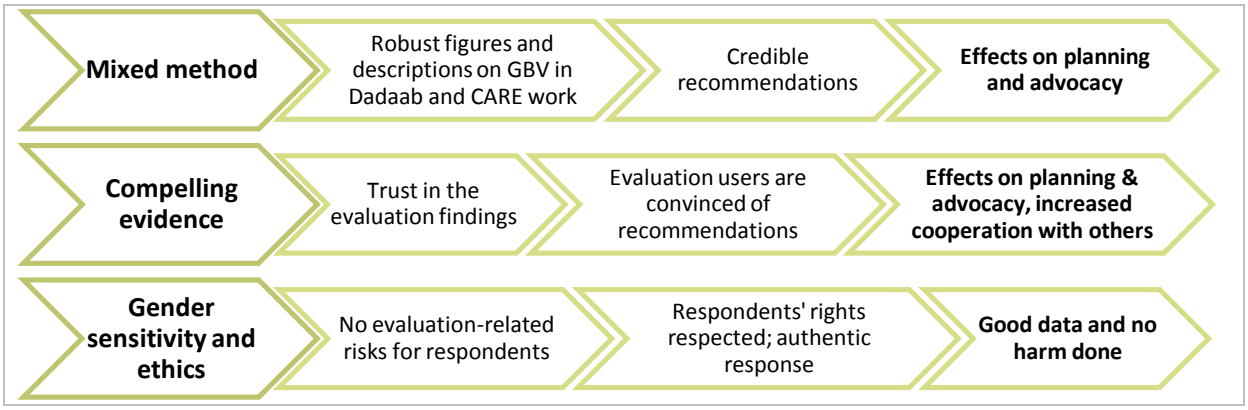


Figure 26: Cause-to-effect chains for Mwangi (2012)

6. Recommendations for effective evaluations

Different paths to effective evaluation

In our analysis above, we identified configurations that led to effective evaluations of interventions on violence against women and girls in development, humanitarian and (post-) conflict contexts. Each configuration or *path* came with its own combination of essential factors leading to success. A factor that is necessary in one effective configuration may be unimportant in a different configuration.

The approaches and methods chosen are only two among several factors that determine whether an evaluation will be useful. Quantitative, qualitative and mixed methods can yield effective evaluations – provided they are combined with the right *conditions*.

➔ ***Recommendation:*** Commissioners should be open to a wide range of approaches and methods, including novel approaches, and encourage evaluators to tailor each evaluation to its specific purpose and context. Both qualitative and quantitative design can lead to effective evaluation.

The diagrams below show the combinations of conditions we have found to make evaluations of interventions on VAWG effective. **Commissioners can use the diagrams to verify whether an evaluation design and context combines all necessary factors.**

The paths in the first diagram have led to effective evaluations regardless of the wider context.

Paths to effective evaluation

A check mark ✓means that the respective condition has to be met; a cross ✗ means the respective condition has to be absent. If both ✓ and ✗ are displayed, it means that the dimension is not important within the individual configuration for the evaluation to be effective.

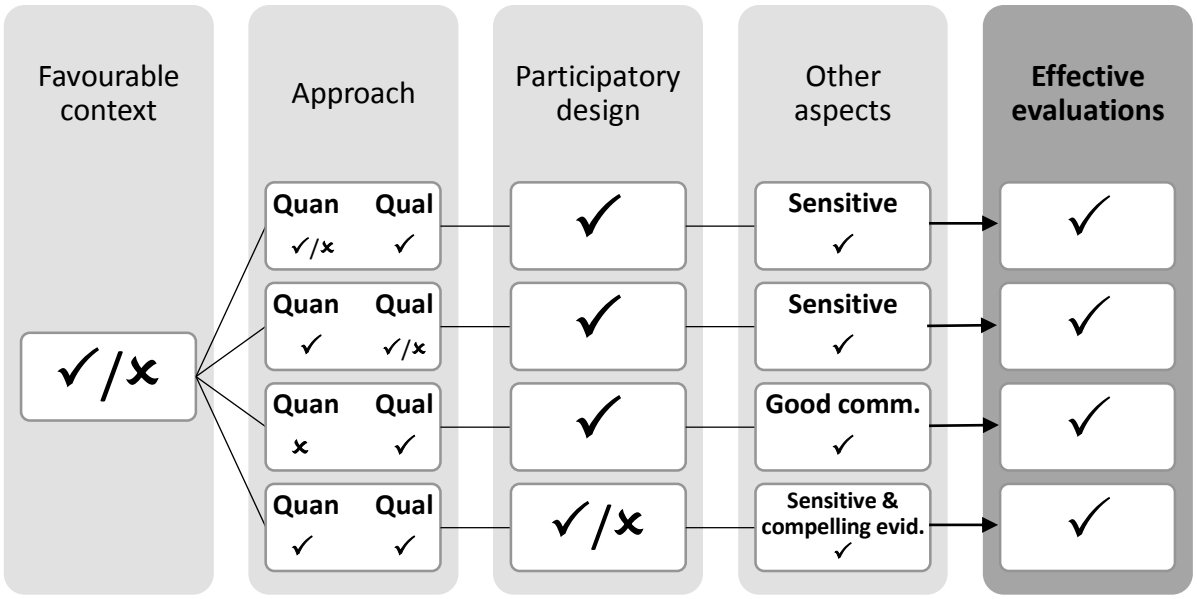


Figure 27: Paths to effective VAWG evaluation, irrespective of context

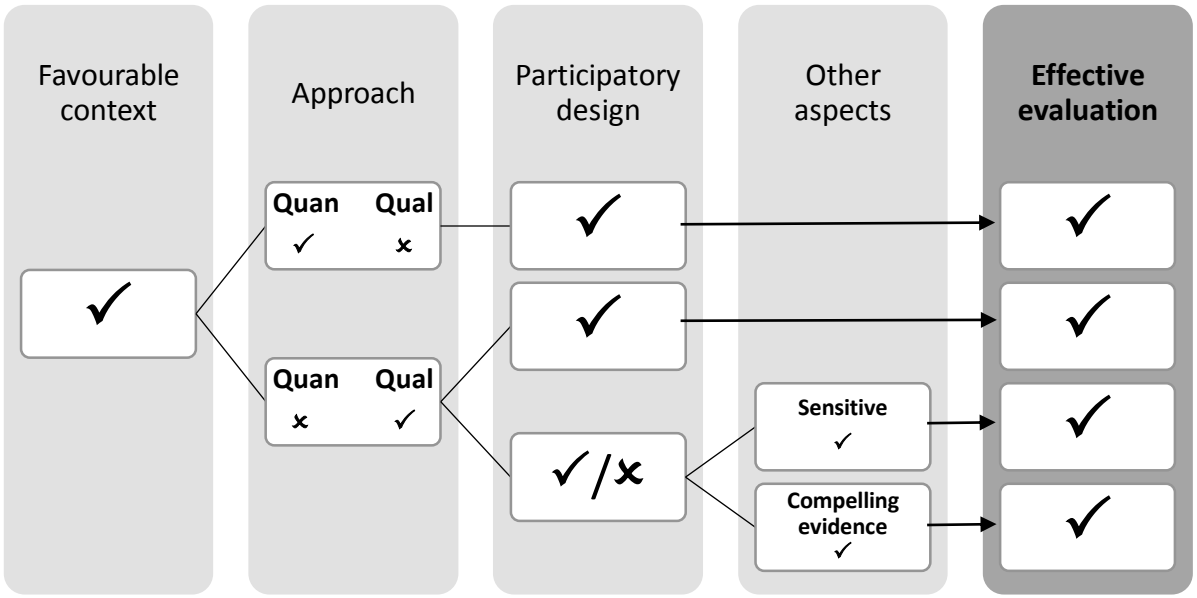


Figure 28: Paths to effective VAWG evaluation in favourable contexts only

The four paths in the second diagram yielded strong positive evaluation effects only in favourable contexts. We have used three dimensions to define favourable context: (i) a stable internal and external environment of the intervention, (iii) relatively simple evaluation tasks and (iii) evaluators with a strong mandate.

Participation and sensitivity – conditions for effective evaluation

Our QCA found two factors - participation and sensitivity (defined in this study as a combination of gender sensitivity and awareness of evaluation-related risks) – to play essential roles in most configurations for effective evaluation of interventions on VAWG.

Participation (defined as the involvement of intervention stakeholders in evaluation design and the interpretation of data) was a *necessary condition* in *paths* to evaluation effectiveness followed by 75% of the evaluations generating strong effects.

➔ ***Recommendation***: Evaluations should be designed and interpreted in consultation with evaluation users to ensure evaluators obtain high quality data, interpret it correctly and produce recommendations that are adapted to the evaluation purpose. Key moments for consultation include:

- TOR development
- Evaluation inception and planning, development of data collection instruments
- Interpretation of findings

Sensitivity. The second *condition* that appeared frequently in *paths* to evaluation effectiveness (82% of evaluations with strong effects) was a combination of gender sensitivity and sensitivity to evaluation-related risks.

➔ ***Recommendation***: Evaluation teams need to be familiar with gender studies, in particular in relation to VAWG. They must observe ethical guidelines, such as the WHO guidelines for research on violence against women and girls (WHO 2001), to prevent violations of the rights of those potentially affected by the evaluation.

Evaluation TOR should refer to specific ethical guidelines and require evaluators to include information on ethical issues in their evaluation or inception report. Evaluation quality assurance mechanisms should monitor gender sensitivity and the observance of ethical guidelines during the evaluation.

Methodological rigour for compelling evidence

Compelling evidence is not the single most important factor for effective evaluation, but it is important in its own right: common sense dictates that accurate data generates better recommendations than faulty information.

Regardless of the expressions evaluation teams may use to describe their methodology (for instance, "rigorous evaluation", "representative survey", "outcome mapping", "operational research", "most significant change method"), any evaluation design includes basic data collection methods. Surveys, focus group discussions, interviews and desk review are the most common building blocks. A straightforward way to verify whether an evaluation design is likely to yield robust findings, is to examine these building blocks.

➔ ***Recommendation***: The inception report or evaluation plan should include a description of the methodology, information on data collection and analysis tools and an assessment of the extent and rigour to which the proposed approach and method can answer the evaluation questions. Abstract terms such as "triangulation" and "participation" should be defined so that readers understand what will happen in the evaluation, how data will be interpreted and what degree of accuracy can be reached.

➔ ***Recommendation***: The methods used for data gathering and analysis should be explained and systematically documented in the evaluation report. Annexes should include the tools used, such as questionnaires and interview guides. Sampling strategies, whether for surveys or for interviews, should be clearly spelled out.

Broader distribution for wider learning

Most of the 39 evaluation reports we examined were well structured and written in an accessible language, but their findings were not systematically disseminated.

➔ ***Recommendation***: Evaluation reports should be published and shared more widely – not only in summaries of key findings. Ideally, they should be shared in full, including the documentation of the methodology, via several channels. Where evaluation participants' rights could be affected by such wide distribution, data should be anonymised.

Annexes

Annex I: Short descriptions of evaluations

The following one-page descriptions of evaluations illustrate the diversity of approaches and methods used in evaluations of interventions on violence against women and girls. These descriptions do not represent findings of our QCA; they serve a purely illustrative purpose.

Our review has found that **methodological choice is only one factor among other conditions that must come together to produce effective evaluation**.

We recommend evaluators, commissioners and other evaluation users **refer to the two decision tree diagrams presented in our recommendations above when reflecting on appropriate evaluation design**.

Zero Tolerance Village Alliance Intervention Model
Final evaluation, 2012, South Africa. Author(s): Craig Carty.

Mixed-method design. Quasi-experiment with pre- & post-test surveys and a control group. Focus group discussions with beneficiaries.

DAC CRITERIA COVERED

☐ Relevance ☐ Effectiveness ☐ Sustainability
☐ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☑ Empowerment
☑ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Impact measurement through pre- and post-test surveys in treatment and control villages allowed to attribute changes to the intervention. Focus group discussions revealed a crucial cause for differences between treatment groups that otherwise would have gone unnoticed.

INTERVENTION EVALUATED

Village intervention programme with four thematic areas: sexual assault, domestic violence, child abuse and HIV/AIDS. Its core was a series of workshops culminating in a ceremony in which villagers pledged support for survivors of gender-based violence (GBV) and people living with HIV.

PURPOSE OF THE EVALUATION

To capture changes in knowledge, attitude and practice with respect to sexual and gender-based violence in intervention villages.

KEY FINDINGS

Voluntary HIV counseling & testing rates, knowledge about post-exposure prophylaxis and services for survivors of GBV increased significantly. Support from village leaders was identified as a crucial factor.

METHODOLOGY

The evaluation used a strong mixed-method design. It consisted mainly of surveys and focus group discussions.

The quantitative part took the form of a quasi-experimental design with pre-test and post-test surveys in two villages where the intervention took place ("treatment") and one control village. The survey that was conducted prior to the intervention served also as a means to identify priority thematic areas of intervention. The endline survey was conducted 12 months after the baseline data was collected.

The evaluators conducted focus group discussions with beneficiaries in villages where the intervention was implemented. The qualitative evidence helped to identify the causes for significant differences between the two treatment villages. The evaluators conclude that village leadership support increases the success of the Zero Tolerance Village Alliance Intervention Model significantly.

In at least one occurrence, qualitative evidence may have been compromised by the interference of a village chief who checked on a focus group discussion conducted multiple times "to ensure that everything was running smoothly". Furthermore, the fact that focus group discussions were not conducted separately for women and men may have affected the women's ability to speak freely about gender-based violence and HIV issues.

The Mehwar Centre – Evaluation of Policies and Procedures
2011, Occupied Palestinian Territory. Author(s): Joanna Creighton and Amer S. Madi.

Comprehensive evaluation of policies and procedures. Qualitative design comprising a literature review and interviews.

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☑ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☐ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Evaluation of policies and procedures that included an assessment of their comprehensiveness, consistency with human rights standards, relevance for the centre's work, as well as their effectiveness and impact.

INTERVENTION EVALUATED

The formulation and implementation of policies and procedures of a center for survivors of gender-based violence in the Occupied Palestinian Territory.

PURPOSE OF THE EVALUATION

Review of the effectiveness of the centre's policies and procedures & their compliance with human rights standards. Broader goal to develop them into a model for other centres that support survivors of gender-based violence.

KEY FINDINGS

Developments of the centre's internal structure and services for survivors of gender-based violence are not fully reflected in its policies and procedures. Priority areas for improvement are policies and procedures related to centre governance and management, case management, outreach and staffing.

METHODOLOGY

Although the evaluation task was limited to the policies and procedures of the centre, the TOR included two distinct sets of questions. To answer these questions, the evaluation used a purely qualitative design.

The first set of questions was concerned with the comprehensiveness of the policies and procedures with regard to the structure and scope of the centre's work, and their consistency with human rights standards. For this set of questions, the evaluation design focused on a review of a wide range of documents, including the centre's documentation of policies and procedures, legal & policy documents of relevant state authorities and literature on gender-based violence in the occupied Palestinian territory.

The second set was concerned with the effectiveness and impact of the centre's policies and procedures. Relevant data was obtained through observation, interviews and workshops with centre staff and beneficiaries and a sample of administrative and case files of the centre.

Both approaches were underpinned by repeated meetings with key stakeholders such as representatives of relevant ministries, of UN Women and the center leadership.

The TOSTAN Programme - Evaluation of Long-term Impact
2008, Senegal. Author(s): Diop, Nafissatou J. et al.

Qualitative component of an impact evaluation. Interviews with stakeholders and beneficiaries.

DAC CRITERIA COVERED

☐ Relevance ☐ Effectiveness ☐ Sustainability
☐ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☑ Changing social norms
☐ Building political will
☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

The qualitative research combined structured observation ("village profiles") with in-depth interviews. Thus the evaluation team was able to interpret interview data in the light of information on the social environment of interviewees.

INTERVENTION EVALUATED

The intervention comprised two types of activities: a health education programme to educate a group of women in each village and a social mobilization strategy that facilitated a public declaration of the villages against female genital mutilation and early marriage.

PURPOSE OF THE EVALUATION

To assess how female genital mutilation and early marriage are understood and dealt with. To assess how the TOSTAN programme had an impact on this.

KEY FINDINGS

Greater support for public declarations and greater awareness of the dangers of female genital mutilation in intervention villages. Neither intervention nor control villages showed a change of opinion on early marriage.

METHODOLOGY

The study is the qualitative component of an impact evaluation. Two different evaluation teams implemented the quantitative and the qualitative component respectively. The qualitative component included field observations and approximately 150 individual interviews that were held in 12 different villages. In ten of these villages the programme was implemented. Selection of the villages was done in accordance with the quantitative research component.

To select interviewees, the evaluation team identified a resource person in each village. The resource person was then tasked to identify women who had participated in the programme, women who had not participated in the programme and male, female and youth leaders. In some cases, administrative personnel who were present during the time of the project implementation were interviewed as well. Furthermore, the evaluators conducted interviews with facilitators who had taught the programme.

Informal interviews were conducted to gather information on the forms of organisation, the actions of committees, the role of women and the situation with respect to female genital mutilation and early marriage in the villages. Those interviews included traditional birth attendants, male head nurses, principals of village schools, teachers and leaders of sports, cultural and religious associations. Based on the interviews, village profiles were developed.

Strengthening Community Safety through Local Government Capacity Building
Final evaluation, 2011, Jamaica. Author(s): Daniel B. Gordon

Participatory project implementation assessment, based on interviews, meetings and focus group discussions.

DAC CRITERIA COVERED

☐ Relevance ☑ Effectiveness ☑ Sustainability
☐ Efficiency ☑ Impact ☑ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☐ Changing social norms
☑ Building political will
☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Participatory assessment of the implementation of a project, based on a scoring system that differentiated between three levels of implementation for each project component. Scores were discussed with project stakeholders throughout the evaluation process.

INTERVENTION EVALUATED

Funding and capacity building for community-based organisations and local government authorities to conduct local community safety audits with the active participation of women in two communities, with a focus on women's safety.

PURPOSE OF THE EVALUATION

To assess effectiveness, impact and sustainability of the project. To assess how effective equality and gender mainstreaming have been incorporated.

KEY FINDINGS

The project was well designed but activities were only partially implemented. No improvement was achieved with regard to the capacity of local authorities to address safety issues.

METHODOLOGY

Participatory assessment of programme implementation, based on a desk review, meetings with stakeholders, focus group discussions with community members and a feedback process with implementing partners and other stakeholders. The feedback process used implementation level scores.

The desk review included the documentation of the project and community profiles. Meetings with project managers, implementing partners and representatives of the central and local governments were held before and during field visits in the communities where the project had been implemented. Furthermore, focus group discussions with community members were conducted in both communities.

The evaluator used a scoring system to assess the level of implementation of different project components. For each component, three levels of implementation were determined: non-implementation, acceptable implementation and ideal implementation.

The implementation scores were discussed with project stakeholders throughout the evaluation process. For the final evaluation report, the component scores were weighted to arrive at an overall score for project implementation.

Prevention of Domestic Violence in Uganda

Final evaluation, 2012, Uganda. Author(s): Danny Harvey et al. Commissioned by Oxfam in Uganda.

Mixed-method design. Survey using cost-saving sampling strategy. Focus group discussions with gender-sensitive design.

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☑ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☑ Changing social norms
☐ Building political will
☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Cost-effective implementation of a survey through Lot Quality Assurance Sampling. The design of data collection tools was informed by careful consideration of research ethics, in particular attention to the safety of survivors of domestic violence.

INTERVENTION EVALUATED

Community-based programme to prevent domestic violence. Activities included community dialogues; the creation of support groups; work with schools, local leaders and opinion makers; as well as the implementation of the "We Can Campaign" (http://uganda.wecanglobal.org/).

PURPOSE OF THE EVALUATION

To assess the relevance, efficiency, effectiveness, impact and sustainability of the programme. To assess the extent to which it was successfully gender mainstreamed.

KEY FINDINGS

Knowledge and awareness of domestic violence increased, acceptance of domestic violence was reduced. However, local coordination mechanisms were not improved. Local leaders proved to be the most important "change makers".

METHODOLOGY

The evaluation used both qualitative and quantitative data collection tools.

To assess changes in knowledge, behaviours and attitudes around domestic violence, a survey was conducted in communities where the programme was implemented. To save costs, Lot Quality Assurance Sampling was used. Due to the lack of a baseline, the evaluators opted to base the measurement of change on recollections of the past by respondents.

The team conducted 18 Focus Group Discussions (FGDs) with a diverse set of active stakeholders and beneficiaries. In some communities, beneficiaries for FGDs were randomly selected among villagers. However, mobilizing community members to participate in FGDs proved difficult.

Furthermore, interviews were held with a wide range of key informants. Some of these interviews were used to collect data for standardized partner assessments. A validation workshop was held towards the end of the evaluation.

The safety of female informants was duly considered: To avoid distress, direct questions about the prevalence of domestic violence were dispensed with; most FGDs were held with women and men separately; locations for FGDs were carefully selected to allow women to express themselves freely.

Intervention with Microfinance for AIDS and Gender Equity
Final evaluation, 2006, South Africa. Author(s): Charlotte Watts et al.

Randomised controlled trial with a threefold sampling strategy and a continuous monitoring of intervention implementation.

DAC CRITERIA COVERED

☐ Relevance ☐ Effectiveness ☐ Sustainability
☐ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☑ Empowerment
☑ Changing social norms
☐ Building political will
☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

The study used an experimental design with a randomised sampling procedure and control villages to attribute impact to the intervention. Qualitative data was gathered throughout the implementation period to monitor the delivery of the programme components.

INTERVENTION EVALUATED

A combined microfinance and gender/HIV training intervention. Under the microfinance component, women received small loans to establish income generating businesses. Under the training component, beneficiaries participated in training on gender, HIV and leadership.

PURPOSE OF THE EVALUATION

To assess the overall impact of the programme and investigate the effects of specific components on the reduction of gender-based violence.

KEY FINDINGS

Compared to women in the control communities, those receiving services in the framework of the IMAGE programme showed a significant reduction in reported levels of physical and sexual intimate partner violence.

METHODOLOGY

The evaluation used an experimental design. Eight villages in a rural province in Southern Africa were pair-matched on estimated size and accessibility. One village from every pair was randomly allocated to receive the intervention.

At the beginning of the programme, quantitative data was gathered from three cohorts: (1) women enrolled in the IMAGE programme and women of the same age from households in control villages who would have been eligible to receive loans, (2) household co-residents of these women aged 14 to 35 years and (3) a random sample of community residents aged 14 to 35 years.

After two years, two sets of interviews were conducted with all cohort (1) individuals who had been eligible at baseline and all cohort (2) individuals who had been successfully interviewed at baseline. After three years, all individuals for cohort (3) who had been eligible at baseline were interviewed.

Interviews were conducted by trained female facilitators in a safe location chosen by the respondents. Interviewers concluded by providing information on local support services.

A qualitative research programme monitored delivery of the intervention. Data was gathered through attendance registers, focus groups, financial monitoring systems, and questions on intervention acceptability.

Capacity Building to Prevent and Respond to Gender-Based Violence

Final evaluation, 2012, Guinea. Author(s): Ashley Jackson.

Mixed-method final evaluation. Baseline data available for one of three components of the intervention.

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☐ Sustainability
☑ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☐ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

The evaluation combines data collected by project partners and newly collected data to provide a holistic assessment of an intervention with several components.

INTERVENTION EVALUATED

Provision of care to survivors of gender-based violence (GBV) perpetrated by government forces who had dispersed a mass rally; capacity building for community-level GBV prevention committees; and training for health care providers on GBV issues.

PURPOSE OF THE EVALUATION

To assess whether the needs of survivors were met and whether communities & health facilities increased their capacities to respond to gender-based violence.

KEY FINDINGS

More survivors served and health care providers trained than planned. Services and trainings appreciated by beneficiaries. Local prevention committees provided guidance that ensured activities met locally felt needs.

METHODOLOGY

The evaluation used different qualitative and quantitative methods for the assessment of the three components of the intervention. Service provision and capacity building components were assessed using end-line data only. For the assessment of the training component, both baseline and end-line data was collected.

Semi-structured interviews were the main source of data. Interviews were conducted with both active stakeholders and beneficiaries. Among the former, project partner staff, health care providers and trainers were interviewed. Among the latter, interviews were conducted with gender-based violence prevention committee members, local leaders, community members who attended prevention activities and survivors of gender-based violence. Interviews with survivors focused on whether and how they had benefitted from the project and how the project could be improved - not the survivors' experiences of gender-based violence. Intake interviews provided additional insights.

For the assessment of the training component, semi-structured interviews were conducted at baseline and endline with facility managers and health care providers in 21 facilities in three regions where the intervention was implemented. Both health care providers who had participated in training as well as well as those who had not received any training were interviewed. Interviews with facility managers formed the basis for facility audits.

The Refentse Model for Post-Rape Care

Formative study, 2009, South Africa. Author(s): Julia C. Kim et al.

Formative, mixed-method design with baseline and endline assessments employing a variety of data collection tools.

DAC CRITERIA COVERED

☐ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☐ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Operational research, with detailed baseline and endline assessments employing both qualitative and quantitative data collection tools. Structured tools such as facility inventory checklists were used to standardise observational data.

INTERVENTION EVALUATED

A model for the integration of nurse-driven, post-rape care into existing reproductive health/HIV services in rural South African hospitals. The model included the establishment of a sexual violence advisory committee, the development of hospital rape management policies, trainings, the introduction of a designated examination room and community awareness campaigns.

PURPOSE OF THE EVALUATION

To develop the model and assess its feasibility and costs. To assess the impact of the intervention on the quality of care delivered.

KEY FINDINGS

It is possible to offer effective post-rape care including post-exposure HIV prophylaxis within rural South African hospitals using existing staff and infrastructure. Nurses can play a central role in this form of care.

METHODOLOGY

Quantitative baseline and end-line data collection was accompanied by on-going observation by a resident evaluation team. The baseline data informed the design of the intervention model.

Both assessments used qualitative and quantitative data collection tools. A facility inventory checklist was used to document the availability of relevant tools & resources and the coordination & roles of service providers. It was verified through individual interviews and walk-through documentation. The quality of clinical care provided to patients was assessed with the help of a structured review of hospital charts.

To investigate the quality of care from the rape survivors' perspective, structured patient interviews were conducted. Interviews were conducted in a private room, in local language. Translations were provided by a female translator.

The baseline assessment included additional data collection, such as key informant interviews and a survey of the knowledge, attitudes and practices of health care workers, social workers and police.

Finally, an economic analysis provided estimates for the additional costs to the health sector of improving post-rape care using the intervention model.

Slavery and Child Labour: Governance and Social Responsibility Project
Mid-term evaluation, 2010, several countries. Author(s): Asmita Naik.

Qualitative mid-term evaluation of a multi-country intervention

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☑ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☐ Changing social norms
☑ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Most conclusions and recommendations were based on information from various sources, including stakeholders and beneficiaries.

INTERVENTION EVALUATED

The intervention comprised three components: research on the psycho-social impact of domestic work on children, a small grant scheme, and advocacy at local, national and international levels. Child domestic workers' advisory committees were created to increase involvement of beneficiaries.

PURPOSE OF THE EVALUATION

To assess the relevance, efficiency, effectiveness, impact, sustainability and replicability of the project and the extent to which it contributed to equality.

KEY FINDINGS

High relevance; impact at international, national and individual levels. Overly complex design (three components, six countries), exceeding the management capacity of implementing partners.

METHODOLOGY

The evaluation employed a purely qualitative design. It comprised a document review, phone interviews with stakeholders at the global level, focus groups discussions with beneficiaries and face-to-face interviews with national stakeholders. The evaluator visited three out of the six countries where the project was implemented. Partner organisations in countries that had not been visited were asked to fill out a self-assessment questionnaire.

The self-assessment questionnaire was the main evaluation instrument. It consisted of open questions grouped around the DAC criteria to be evaluated. According to the evaluator, it also served as a basis for developing guidelines for interviews and focus group discussions.

Focus group discussions were conducted with child domestic workers, parents and community members in each of the countries visited. Among national stakeholders interviewed for the evaluation were staff of implementing partners, representatives of other national civil society groups, local representatives of international organisations and government officials.

For focus group discussions and interviews with local stakeholders, an independent translator accompanied the evaluator. Another independent consultant was hired by the evaluator to peer-review the draft report.

Ending Domestic Violence in Rwanda
Mid-term evaluation, 2011, Rwanda.
Author(s): Dorothy Omollo-Odhiambo and Tom Odhiambo.

Cross-sectional survey, focus group discussions and interviews used for impact measurement of a complex intervention.

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☑ Empowerment
☑ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

A survey based on a clustered random sampling procedure was used for a rough yet rigorous impact assessment. Focus group discussions and interviews aided interpretation of the survey data.

INTERVENTION EVALUATED

Funding and capacity building for national civil society organisations that work towards ending domestic violence in Rwanda. Supported programmes included psychosocial services, awareness raising, community empowerment, collaboration with local authorities & economic empowerment.

PURPOSE OF THE EVALUATION

To evaluate the strategies and approaches employed by implementing organisations in their work against domestic violence.

KEY FINDINGS

The chosen strategies proved successful. Almost half of the target population had received information on domestic violence and/or counselling services through the programmes of implementing organisations.

METHODOLOGY

The evaluation used a mixed-method design, comprising a survey, focus group discussions and key informant interviews. The survey was carried out with a representative sample of direct and indirect beneficiaries of all implementing organisations. In this way, quantitative data for impact measurement was available for all the interventions by implementing organisations. The sampling was randomised and used clustering.

The implementing organisations who were supported by the "Ending Domestic Violence Project" used diverse strategies and approaches in their work. As the same questionnaire was used for the survey across programme areas, it could not measure impact based on the logic of the implementing organisations' individual programmes. Instead, it focused on impact as defined by the "Ending Domestic Violence Project". Respondents were asked whether knowledge, attitudes and practices related to domestic violence had changed due to interventions by implementing organisations, based on their self-assessments.

While the interviews for the survey were conducted by research assistants, the lead evaluators gathered qualitative data. They interviewed key informants and held focus group discussions with beneficiaries. The qualitative data was mainly used to provide background for the interpretation of the quantitative data.

Assessment of the 2nd Phase of the 'We Can' Campaign in India
Mid-term evaluation, 2010, India.
Author(s): Anuradha Rajan and Swati Chakraborty

Mid-term evaluation that combines a quantitative assessment of impact with a thorough analysis of qualitative data.

DAC CRITERIA COVERED

☐ Relevance ☐ Effectiveness ☐ Sustainability
☐ Efficiency ☑ Impact ☑ Other (see "purpose")

DFID PRIORITIES

☐ Empowerment
☑ Changing social norms
☐ Building political will
☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Critical reflection on what impact means in the framework of the evaluated intervention. Advocating for a nuanced understanding of the concept of change based on a thorough analysis of qualitative data. Development of an innovative categorization of "change makers".

INTERVENTION EVALUATED

The 'We Can' campaign works with individuals, so called "change makers" (CM), who commit to rejecting violence against women. In the 2nd phase of 'We Can', the focus of the campaign was on re-engaging CMs who were fostering change in their communities and recruited at an earlier date.

PURPOSE OF THE EVALUATION

To assess the impact of the campaign on individual CMs and the communities they live in. To establish a framework for ongoing monitoring.

KEY FINDINGS

60% of CMs showed a deepened understanding of the campaign issues and continued to support the campaign through activities. 70% of CMs "circles of influence" reported personal change.

METHODOLOGY

The evaluation assessed impact on two levels: The CMs and their "circle of influence". Assessments were carried out at four study sites in three states in India that were purposively selected to ensure differences between major intervention areas were covered.

CMs were interviewed with the help of a structured interview guide to assess the impact the campaign had on them after their recruitment. They were also invited to a one-day workshop where the evaluators explored key phases of the change makers' lives around the time the CM had joined the campaign. Women and men in the change makers' social environment ("circle of influence") were asked in semi-structured interviews and focus group discussions about the impact the CMs had achieved.

CMs were randomly selected from lists provided by implementing partners. Persons in the change makers' "circles of influence" were identified by the CM. They were asked to identify those persons in their environment they believe the campaign had an impact on.

The evaluators used the qualitative data obtained in workshops, interviews and in focus group discussions to develop a categorisation of CM that takes into account the change makers' personal circumstances. They strongly advocated for a recognition of the different meanings CMs give to the concept of change.

Reducing Violence against Women & Enhancing Access to Justice for Women in Humanitarian Emergencies & Conflict Areas
Final evaluation, 2012, four African countries. Author(s): Althea Rivas.

Evaluation of a complex, multi-country intervention with a purely qualitative design.

DAC CRITERIA COVERED

☑ Relevance ☑ Effectiveness ☑ Sustainability
☑ Efficiency ☑ Impact ☐ Other (see "purpose")

DFID PRIORITIES

☑ Empowerment
☑ Changing social norms
☐ Building political will
☑ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Faced with the task to evaluate a very complex multi-country intervention in a short time frame, the evaluator used every opportunity to collect additional data. The evaluation report is clearly structured and succinct.

INTERVENTION EVALUATED

Financial and capacity building support to civil society organisations in four African countries for programmes to reduce VAW and enhancing women's access to justice. Activities included the establishment of support groups, advocacy, psycho-social services, economic empowerment and training for civil servants.

PURPOSE OF THE EVALUATION

To assess the quality of the design and implementation of the supported programmes. To assess coordination and cooperation among the project partners.

KEY FINDINGS

Strategic direction and monitoring of the programmes by the leading project partner should be strengthened to develop more targeted activities and clear objectives. Delays in implementation; geographic scope too wide.

METHODOLOGY

The evaluation employed a qualitative design. It included a desk-based review of project documentation and relevant laws and policies, and a financial analysis. Interviews were held with a range of stakeholders in each of the intervention countries, as well as focus group discussions with beneficiaries.

In addition to collecting a common set of comparable data in all programme countries, the evaluator used specific data collection tools in individual countries. This included direct observation, a stakeholder workshop and collection of case studies of survivors of gender-based violence.

The evaluation faced multiple challenges: A baseline assessment was not available. Approximately 15 days were reserved for data collection for a complex, multi-country intervention. One of the countries in which the project was implemented could not be visited, thus focus group discussions with beneficiaries could not be conducted. Some of the activities were completed several years prior to the evaluation and responsible staff members had moved on to other positions.

The evaluator used a grading system based on the DAC criteria to summarise findings and focused recommendations on planning, coordination and cooperation among the project partners.

Review of the 'We Can' Campaign in Uganda

Mid-term evaluation, 2012, Uganda. Author(s): Joseph Rujumba et al. Commissioned by Oxfam in Uganda.

Baseline assessment with a predominantly quantitative design.

DAC CRITERIA COVERED

☐ Relevance	☑ Effectiveness	☐ Sustainability
☑ Efficiency	☐ Impact	☑ Other (see "purpose")

DFID PRIORITIES

- ☐ Empowerment
- ☑ Changing social norms
- ☐ Building political will
- ☐ Providing services

COMMENDABLE ASPECTS OF THE EVALUATION

Baseline assessment provided rigorous evidence for key performance indicators. Qualitative data was used to generate additional insights into the causes, forms and effects of gender-based violence as well as suggestions for improving the effectiveness of the intervention.

INTERVENTION EVALUATED

The 'We Can' campaign worked with local partners to encourage women and men, girls and boys to become "change makers", who publicly commit to rejecting violence against women. Activities included the formation of a campaign alliance, development of communication materials, the formation of clubs at universities, work with religious leaders and awareness raising rallies.

PURPOSE OF THE EVALUATION

To assess the efficiency and effectiveness of the campaign implementation. To identify baseline information for key performance indicators.

KEY FINDINGS

A third of the target population was aware of the campaign and 7% of respondents reported that they were change makers. 43% of change makers were male, 57% female.

METHODOLOGY

The evaluation adopted a predominantly quantitative design. A household survey was implemented in seven of the twelve districts that were covered by the campaign. Interviews for the survey were semi-structured. The survey was used to assess baseline information on knowledge, awareness, attitudes and tolerance of domestic violence as well as awareness of and participation in the 'We Can' campaign.

Qualitative data provided additional evidence. Survivors of gender-based violence, who were identified through the survey, were interviewed in-depth to generate a more nuanced understanding of the causes of gender-based violence, access to services and the changes observed and desired in the communities.

The assessment of the state of campaign implementation was largely based on a literature review and interviews with Oxfam staff, district officials and other key informants. Focus group discussions with community leaders, change makers and other community members were used to triangulate this information. In workshops, members of district alliances were invited to provide feedback and generate suggestions for improving the effectiveness of the remaining phase of the campaign.

Annex II: Methodological notes

Review team

The core Review team was made up of two independent consultants, **Michaela Raab** and **Wolfgang Stuppert**. Michaela is an evaluator and gender justice specialist with more than two decades of experience in world-wide development, humanitarian and peace building work. Wolfgang is a PhD candidate in social sciences who has carried out empirical research on civic activism, civil society development and democratization, using both qualitative and quantitative methods. The core team was assisted by **Miruna Bucurescu, Scout Burghardt, Sanja Kruse, Astrid Matten** and **Paula Pustulka**, who performed substantive coding tasks and transcribed interviews. All coders hold M.A. degrees and have a social studies background, in particular in gender studies and qualitative research. An external specialist in QCA, **Julian Brückner** (researcher at the Berlin Social Science Center), lent his support as a sounding board on QCA. This included input to a specialised article presenting the Review as a case study for the use of QCA.

Quality assurance

Several internal and external mechanisms were used to ensure a rigorous review process.

Internal mechanisms

The mix of data collection tools and triangulation of responses to our survey and in process tracing interviews warranted high data quality. All data collection tools were pre-tested. Coders were trained in two half-day sessions; their work was regularly monitored.

Questions related to QCA were discussed with academic QCA specialist Julian Brückner at key phases of the Review. The Review team documented the phases of the Review in a Scoping Report, an Inception Report and a dedicated blog www.evawreview.de.

External monitoring

At the beginning of the Review, DFID established an external **Reference Group (RG)** composed of specialists in evaluation and VAWG: Joelle Barbot, Krishna Belbase, Valeria Carou-Jones, Katie Chapman, Sabrina Evangelista, Jennifer Leith, Helen Lindley, Clare McCrum, Judith McFarlane, Jodi Nelson, Fiona Power, Amanda Sim, Inga Sniukaite, Zoe Stephenson and Jeanne Ward. The deliverables generated throughout the Review – a tentative model presenting conditions for evaluation effectiveness, as well as drafts of the Scoping, Inception and Review Reports – were reviewed by DFID and the RG. Feed-back was incorporated in subsequent phases of the Review.

Furthermore, the Scoping and Inception Reports were examined by the **Specialised Evaluation and Quality Assurance Service** (SEQAS) with particular attention to the QCA methodology. SEQAS comments and the Review team's response were appended to the final Inception Report. Finally, the Review blog www.evawreview.de received comments and suggestions which we used in our research.

How QCA Works

QCA is based on the assumption that several cause-to-effect chains coexist for the same effects. It examines sets of ***conditions*** in relation to specific ***outcomes***. In our Review, the approaches and methods, the context in which evaluations take place and the adherence to quality standards, are *conditions.* Evaluation effects are the *outcome* that we have studied.

Every case – i.e. every evaluation in our QCA analysis – can be described as a configuration of *conditions*. If two cases share the same *outcome* yet differ in one *condition*, then we can conclude that this *condition* is not *necessary*. It can be removed from the configuration that leads to the *outcome*. By systematically identifying and eliminating such redundant *conditions*, QCA identifies basic configurations (= *paths*) that lead to the outcome.

QCA differentiates between ***necessary*** and ***sufficient conditions***:

- *Necessary conditions* must be present for the outcome to exist. But the presence of a *necessary condition* does not mean the *outcome* always occurs: sometimes, other *conditions* are needed as well to produce the *outcome*. Yet, the *necessary condition* has to be part of all configurations of *conditions* that lead to the *outcome*.
- If *sufficient conditions* are present, then the *outcome* exists; i.e. a *sufficient condition* alone can cause the *outcome*. But a *sufficient condition* does not have to be present for an *outcome* to occur; the *outcome* can also be caused by other *conditions* (or combinations of conditions) without that particular *sufficient condition*.
- What holds true for individual *conditions* also applies to combinations of *conditions*: They can be *necessary* and/or *sufficient* for the *outcome* to occur.

QCA helps identifying ***paths*** that lead to effective evaluations. A *path* is a sufficient combination of conditions. The group of *paths* identified for a set of cases is called a ***solution*** in QCA. In this Review it describes all evaluations that have shown a combination of conditions to be effective.

We have worked with two types of solutions, i.e. sets of *paths* to effective evaluation: ***complex*** and ***intermediate solutions***. In *complex solutions*, QCA uses only configurations that have been observed in the actual cases analysed. In *intermediate* solutions, the researcher 'tells' QCA to make simplifying assumptions. For example, we can instruct QCA to assume that the presence of a certain *condition* will lead to the *outcome*. Based on that, QCA takes into account additional, non-observed configurations. The result of such an operation is a *solution* which includes fewer *paths* than a *complex solution*. Such a solution is more easily interpretable.

Distribution of sub-dimensions for composite conditions

With the exception of "participatory design", all conditions in our QCA are composed of several dimensions of evaluation practice. "Favourable context", for example, consists of the dimensions "conducive circumstances", "simplicity of evaluation task" and "powerful mandate".

To determine the presence of each of the six composite conditions, we first assessed the degree to which each sub-dimension was present or not. Then we aggregated these values to arrive at an overall value for the respective condition.

Depending on how we assumed the sub-dimensions to relate to each other, we used different aggregation rules. For instance, we considered both clear presentation and wide distribution to be necessary elements of good communication. Therefore, we took the minimum value of the two sub-dimensions as the value for "good communication". That is, if either clear presentation or wide distribution was absent, we deemed good communication to be absent.

The tables below show the percentage to which the sub-dimensions of a condition were present or absent when we assessed the respective condition. For instance, where "favourable context" was deemed to be present, "conducive circumstances" was present in 52% of the cases and "powerful mandate" was present in 100% of the cases.

Favourable context

	Conducive circumstances	*Simplicity of evaluation task*	*Powerful mandate*
*Favourable context **present***	 0% 100%	 0% 100%	 0% 100%
*Favourable context **absent***	 0% 100%	 0% 100%	 0% 100%

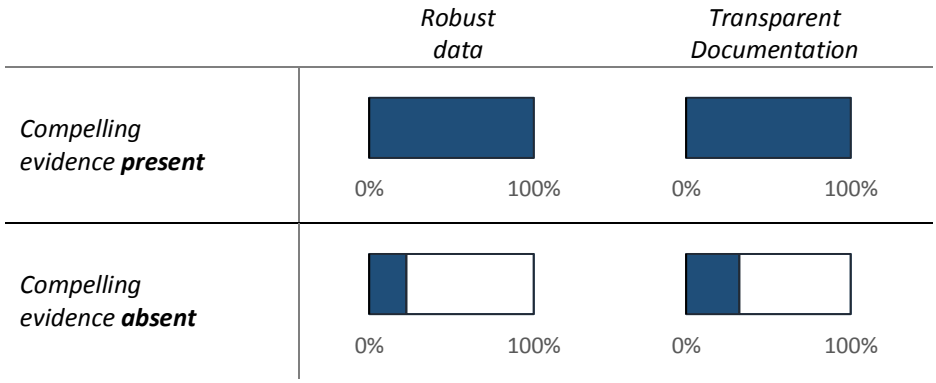
Good communication

	Presentation	*Distribution*
*Good communication **present***	 0% 100%	 0% 100%
*Good communication **absent***	 0% 100%	 0% 100%

Sensitivity to GBV

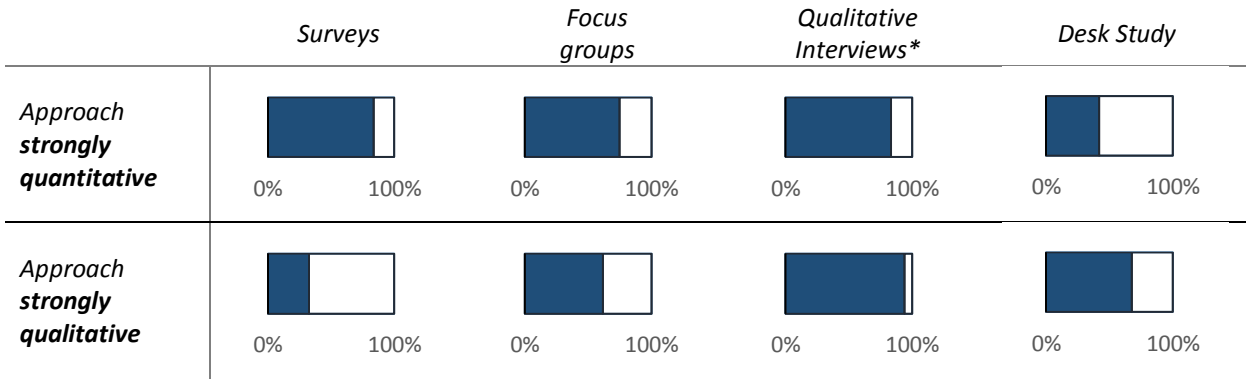
	Sensitivity to eval.-related risks	*Sensitivity to prof. context*
*Sensitivity to GBV context **present***	 0% 100%	 0% 100%
*Sensitivity to GBV context **absent***	 0% 100%	 0% 100%

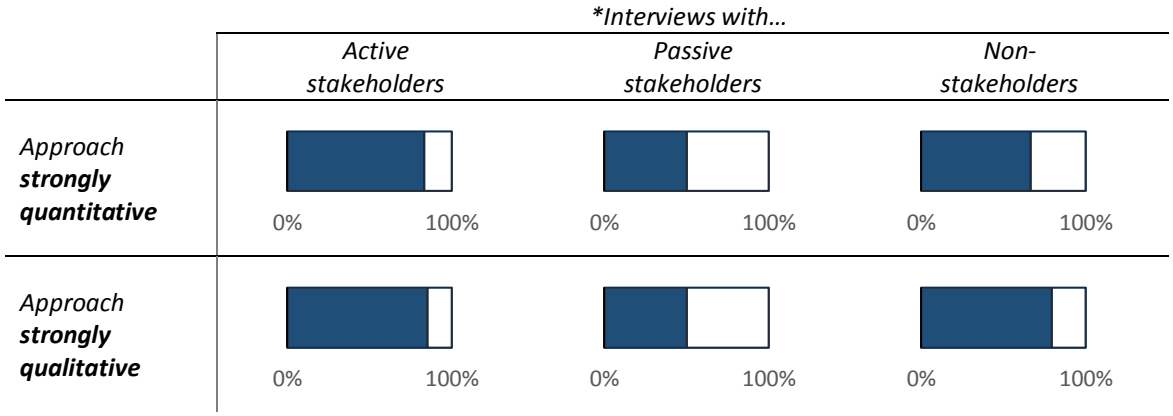
Compelling evidence



Approach

The basis for the assessment whether an approach was strongly quantitative or strongly qualitative was not just the type of methods used, but also to which extent the evaluators relied on the data obtained through those approaches when drawing their conclusions. As individual conclusions can be backed both by qualitative and quantitative data, the approach of an evaluation can be strongly qualitative and strongly quantitative at the same time. The table below, therefore, does not show how we have aggregated the two conditions. It indicates how prevalent specific methods have been among the evaluations whose approach was assessed as strongly qualitative or strongly quantitative.





Annex III: List of evaluations used in QCA

Author[16]	Year	Title	Commissioning organisation
CARE (Mekonnen, Alemayehu)	2009	Health Improvement and Women Owned Transformation (HIWOT) Project End line Evaluation Report	CARE
Carty, Craig	2012	Zero Tolerance Village Alliance Intervention Model	Population Council
Chibuta, Juliet Kaira	2011	Final Project Evaluation Access to Justice for Refugee Women and Girls in Tanzania	One World Trust
Creighton, Joanne et al.	2011	The Mehwar Centre Evaluation of policies and procedures	UN Women
De Boodt, Kristien	2009	Final Evaluation Abatangamuco	CARE
Diop, Nafissatou et al.	2008	Long-term Impact of the TOSTAN Programme on the Abandonment of FGM/C and Early Marriage: Results from a qualitative study in Senegal	Population Council
Drinkwater, Michael J.	2012	Addressing the Heart of the Matter	COVAW Bangladesh
Elmqvist, Madeleine	2011	Review of Kvinna till Kvinna's Georgia Programme 2007-2011	Kvinna till Kvinna
Family Support Institute (Khasiani, Shanyisa)	2012	Tostan Pilot Project on "Ending FGM/C" in Northwest Zone and Northeast Zone of Somalia	UNICEF
Fawzi El-Solh, Camillia and Michael Bernhardt	2011	Act to End Violence against Women Iraq Project	UN Women
Germann, Dorsi and Elias Zedan	2010	Evaluation of "Combating Domestic Violence" Program of TRUST, West Bank, Palestine	Caritas Germany
Gordon, Daniel B.	2011	Strengthening Community Safety through Local Government Capacity Building	UNDP Jamaica
Hailu, Yewubdar	2010	Securing the Future of Afar Pastoralist Women through Ending Female Genital Mutilation	Development Fund Norway
Harvey, Danny and Mary Ssonko	2012	End of Project Evaluation Final Report Oxfam GB in Uganda Prevention of Domestic Violence Project	Oxfam GB
Ingdal, Nora et al.	2008	Mid-Term Review of Project Practice Reduction and Awareness on Female Genital Mutilation (FGM)	YWCA Global
Kim, Julia C. et al.	2009	The Refentse Model for Post-Rape Care: Strengthening Sexual Assault Care and HIV Post-Exposure Prophylaxis in a District Hospital in Rural South Africa	Population Council

[16] Several reports have been co-authored by several specialists. The authors listed here are the ones who appear in the first (or first and second) place on the evaluation report.

Author[16]	Year	Title	Commissioning organisation
Kuneviciute, Ieva	2012	Anti-Trafficking Campaign in Kosovo – Final Evaluation Report	Terre des hommes
Lepetit, Patricia Garcia ; Michelen Ortolá, Marta	2012	Final Evaluation Report Victims' Shuras: Women Victims of War mobilising towards reconciliation and justice	medica mondiale
Marrar, Shuaa	2010	Evaluation Report UNIFEM occupied Palestinian territory Sabaya Programme	UNIFEM
Moen, Hanne Lotte et al.	2012	A comparative evaluation of Fokus FGM projects in East Africa	FOKUS
Mwangi, Gladys Kabura	2012	GBV Program Evaluation Kenya	CARE
Naik, Asmita	2012	Independent Evaluation of United Nations Inter-Agency Project on Human Trafficking in the Greater Mekong Sub-Region Phase III	United Nations Inter-Agency
Naik, Asmita	2010	Slavery and child labour: governance and social responsibility project	Anti-Slavery International
Odhiambo, Tom and Dorothy Omollo-Odhiambo	2011	Mid-Term Evaluation of the Ending Domestic Violence Project in Rwanda	Norwegian People's Aid
Orgocka, Aida and Nikolina Kenig	2012	Final Evaluation of the UN Joint Programme Strengthening National Capacities to Prevent Domestic Violence	UNDP
Pittman, Alexandra	2010	Making Gender Based Violence Programming Explicit: A Review	Oxfam International
Raab, Michaela	2011	Strategic Review of the Coalition for Women's Human Rights in Conflict	Rights and Democracy
Rajan, Anuradha	2010	Assessment of We Can Phase II India Report	Oxfam GB
Rivas, Althea	2012	End of Project Evaluation: Reducing Violence against Women and Enhancing Access to Justice in Humanitarian Emergencies and Conflict Areas in Sierra Leone, Burundi, Democratic Republic of Congo and Uganda	Action Aid International
Robinson, Victor C.	2011	Putting the Jigsaw Together – CARE International Sri Lanka's Violence against Women Intervention in Batticaloa	CARE International
Robinson, Victor C.	2012	COVAW Project Final Evaluation	COVAW Bangladesh
Rujumba, Joseph et al.	2012	Midterm Review of the We Can Campaign to End All Violence against Women, Oxfam GB Uganda	Oxfam GB
Shaheed, Aisha Lee	2011	Transnational Responses to Violence Against Women in the Name of 'Culture'	Women Living under Muslim Law

Author[16]	Year	Title	Commissioning organisation
Smith, Janel et al.	2012	Evaluating Effectiveness of the Clinical Care for Sexual Assault Survivors Multimedia Training Tool in Humanitarian Settings	International Rescue Committee (IRC)
Sotirovic, Vilana Pilinkaite	2012	Final Evaluation Report for the UNDP Project Combating Sexual and Gender Based Violence	UNDP
Townsend, Stephanie and Heimburger, Angela	2010	Evaluation of the Sexual Violence Research Initiative 2010	Global Forum for Health Research
Turnbull, Beverley	2011	Independent Evaluation Report, Pacific Prevention of Domestic Violence Program	New Zealand Aid
UN Women (Gyalang, Nirmal)	2012	Partnership for Equality and Capacity Enhancement (PEACE): Towards Implementation of UNSCRs 1325 und 1820 Project	UN Women
Villavicencio S., Rosa	2012	Final Evaluation Big Lottery Fund Grant – Violence against women in Peru: Improving health and promoting rights	Womankind

Annex IV: Other literature used

Ajema, Carolyne and Buluma Bwire (2009): Standards required in maintaining the chain of evidence in the context of post rape care services: Findings of a study conducted in Kenya

Amo, Courtney; Cousins, J. Bradley (2007): Going through the process: An examination of the operationalization of process use in empirical research on evaluation. In *New Directions for Evaluation* 2007 (116), pp. 5–26. DOI: 10.1002/ev.240.

AusAID (2008): Violence against Women in Melanesia and East Timor. Available online http://www.ode.ausaid.gov.au/publications/Documents/vaw_cs_full_report.pdf

Barker, Gary; Ricardo, Christine and Nascimento, Marcos (2007): Engaging men and boys in changing gender-based inequity in health: Evidence from programme interventions. Published by World Health Organisation. Available online http://www.who.int/gender/documents/Engaging_men_boys.pdf.

Barsoum, Ghada et al (2011): National Efforts toward FGM-free Villages in Egypt: The Evidence of Impact. Working Paper. Unter Mitarbeit von Nadia Rifaat, Omaima El-Gibaly, Nihal Elwan, Natalie Forcier. Published by Population Council. New York, New York (Working Paper No.22).

Bass, Judith K.; Annan, Jeannie; McIvor Murray, Sarah; Kaysen, Debra; Griffiths, Shelly; Cetinoglu, Talita et al. (2013): Controlled Trial of Psychotherapy for Congolese Survivors of Sexual Violence. In: *N Engl J Med* 368 (23), S. 2182–2191. DOI: 10.1056/NEJMoa1211853.

Batliwala, Srilatha; Pittman, Alexandra. (2010): Capturing Change in Women's Realities. A Critical Overview of Current Monitoring & Evaluation Frameworks and Approaches. Edited by Association for Women's Rights in Development (AWID). Toronto, Canada. Available online at http://www.awid.org/About-AWID/AWID-News/Capturing-Change-in-Women-s-Realities, checked on 9/25/2013.

Beach, Derek; Pedersen, Rasmus Brun (2013): Process-tracing methods. Foundations and guidelines. Ann Arbor: University of Michigan Press.

Beardon, Hannah (2013): From paper rights to living rights: RHV in the Gambia. Case study by Hannah Beardon. Published by Oxfam GB. Available online http://api.ning.com/files/bIBgOihCJLCdzjdD*nDMgCoajmW7hqsaLLIq*hAUYMiwjQ-hT5dbtU6RnNcIwfmwjkw42aVLwt32Ef-U0EiSnTN6C*-38CQf/4.RHVTheGambiacasestudy.pdf.

Blanc, Anne K., Melnikas, Chau and Stoner (2013): A review of the evidence on multi-sectoral interventions to reduce violence against adolescent girls. Published by Girleffect.org.

Bloom, S. (2008): Violence against women and girls, a compendium of monitoring and evaluation indicators. USAID. Washington. Available online at http://www.cpc.unc.edu/measure/tools/gender/violence-against-women-and-girls-compendium-of-indicators.

Befani, Barbara, Simone Ledermann, and Fritz Sager (2007). Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application. Evaluation 13: 171-192.

Befani, Barbara (2013). Between complexity and generalization: Addressing evaluation challenges with QCA. Evaluation 19: 269-283.

Clifton, Deborah (2012): Gender Equality in the East Africa Crisis Response. Published by Disasters Emergency Committee.

DAC Guidelines and Reference Series Quality Standards for Development Evaluation (2010). [S.l.]: OECD Publishing.

Danida evaluation guidelines (2012). Copenhagen: Ministry of Foreign Affairs of Denmark.

DFID Evaluation Department: Quality assurance template, Entry level evaluation product. Edited by DFID Evaluation Department.

DFID Evaluation Department: Quality Assurance Template, Exit level evaluation product. With assistance of DFID Evaluation Department.

Diana J. Arango, Mary Ellsberg, Matthew Morton, Floriza Gennari, Sveinung Kiplesund: Interventions to prevent or reduce violence against women and girls: a systematic review of reviews. Available online http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42013004422.

Ellsberg, Mary Carroll; Heise, Lori (2005): Researching violence against women. A practical guide for researchers and activists. Washington, D.C: World Health Organisation.

Feinstein, O. N. (2002): Use of Evaluations and the Evaluation of their Use. In *Evaluation* 8 (4), pp. 433–439. DOI: 10.1177/13563890260620621.

Francisco, L.; Abramsky, T.; Kiss, L.; Michau, L.; Musuya, T.; Kerrigan, D. et al. (2013): Violence Against Women and HIV Risk Behaviors in Kampala, Uganda: Baseline Findings from the SASA! Study. In: *Violence Against Women* 19 (7), S. 814–832. DOI: 10.1177/1077801213497557.

Gage, A.; Dunn, M. (2010): Monitoring and Evaluating Gender-Based Violence Prevention and Mitigation Programs - A Facilitator's Training Guide. Carolina Population Center. Available online http://www.igwg.org/igwg_media/gbv-monitor-eval/gbv-me-facilitators-guide.pdf.

Goertz, Gary; Mahoney, James (2012): A tale of two cultures. Qualitative and quantitative research in the social sciences. Princeton, N.J: Princeton University Press.

Greene, J. G. (1988): Stakeholder Participation and Utilization in Program Evaluation. In *Evaluation Review* 12 (2), pp. 91–116. DOI: 10.1177/0193841X8801200201.

Heise, L. (2011): What works to prevent partner violence: an evidence overview. Published by STRIVE Research Consortium.

Hustache, Sarah; Moro, Marie-Rose; Roptin, Jacky; Souza, Renato; Gansou, Grégoire; Mbemba, Alain et al. (2009): Evaluation of psychological support for victims of sexual violence in a conflict setting: results from Brazzaville, Congo. In: *Int J Ment Health Syst* 3 (1), S. 7. DOI: 10.1186/1752-4458-3-7.

Keesbury, Jill and Ian Askew (2010): Comprehensive responses to gender-based violence in low-resource settings: Lessons learned from implementation. Lusaka, Zambia.

Keesbury, Jill and Mary Zama, Sudha Shreeniwas (2009): The Copperbelt model of integrated care for survivors of rape and defilement: Testing the feasibility of police provision of emergency contraceptive pills.

Kim, J. C.; Askew, I.; Muvhango, L.; Dwane, N.; Abramsky, T.; Jan, S. et al. (2009): Comprehensive care and HIV prophylaxis after sexual assault in rural South Africa: the Refentse intervention study 338 (mar13 1), S. b515. DOI: 10.1136/bmj.b515.

Kirkhart, Karen E. (2000): Reconceptualizing evaluation use: An integrated theory of influence. In *New Directions for Evaluation* (88), pp. 5–23.

Konigson, Asa (2012): Review of Kvinna till Kvinna's Georgia Programme 2007-2011.

Ledermann, Simone (2012) "Exploring the Necessary Conditions for Evaluation Use in Program Change", *American Journal of Evaluation* 33(2): 159-178.

Mahoney, James (2012): The Logic of Process Tracing Tests in the Social Sciences. In *Sociological Methods & Research* 41 (4), pp. 570–597. DOI: 10.1177/0049124112437709.

Mai, L. T. P. (2012): Community response to Domestic Violence (DV): lessons learnt from interventions in Vietnam and policy suggestions (Supplement 1). Available online http://www.safetylit.org/citations/index.php?fuseaction=citations.viewdetails&citationIds=citjournalarticle_382325_3&sha=1.

Manuate, Carmen (2013): RAISING HER VOICE, breaking the silence. RHV Honduras A Case Study by Carmen Manuate.

McAslan Fraser, Erika (2011): Use of participatory methods in VAWG evaluations. GSDRC (GSDRC Helpdesk Research Report) 774.

McIoughlin, Claire (2011): Impact evaluations of programmes to prevent and respond to violence against women and girls. GSDRC (GSDRC Helpdesk Research Report, 789). Available online at http://www.gsdrc.org/docs/open/HD789.pdf.

Ministry of Finance and Economic Development and UNICEF in Ethiopia: Progress in Abandoning Female Genital Mutilation/ Cutting and Child Marriage in Self-Declared Woredas. Available online http://www.unicef.org/evaldatabase/files/Ethiopia_FGM_Final.pdf.

Morrison, A.; Ellsberg, M. Bott S. (2007): Addressing gender-based violence: a critical review of interventions. In: *The World Bank Observer* 22 (1), S. 25–51. Available online http://wbro.oxfordjournals.org/cgi/reprint/22/1/25.pdf.

Morrison, Andrew; Ellsberg, Mary and Bott, Sarah (2004): Addressing Gender-Based Violence in the Latin American and Caribbean Region: A Critical Review of Interventions. World Bank Policy Research Working Paper 3438

Njuki, Rebecca; Okal, Jerry; Warren, Charlotte E.; Obare, Francis; Abuya, Timothy; Kanya, Lucy et al. (2012): Exploring the effectiveness of the output-based aid voucher program to increase uptake of gender-based violence recovery services in Kenya: A qualitative evaluation. In: *BMC Public Health* 12 (1), S. 426. DOI: 10.1186/1471-2458-12-426.

OECD: Evaluating development activities - 12 Lessons from the OECD DAC. Providing evidence on results for learning and decision making. Available online at http://www.oecd.org/dac/peer-reviews/12%20Less%20eval%20web%20pdf.pdf, checked on 9/30/2013.

Orrnert, A. (2012): Evaluations of Programmes Related to Violence against Women and Girls (GSDRC Helpdesk Research Report). Published by the Governance and Social Development Resource Centre, University of Birmingham. Birmingham.

Patton, Michael Quinn (2008): Utilization-focused Evaluation. 4th ed. Thousand Oaks: SAGE Publications.

Ragin, Charles C. (1987): The comparative method. Moving beyond qualitative and quantitative strategies. Berkeley, Calif.: University of California Press. Available online at http://books.google.com/books?id=mZi17vherScC.

Ragin, Charles C. (2008): User's guide to Fuzzy-Set/Qualitative Comparative Analysis. University of Arizona. Tucson. Available online at http://www.u.arizona.edu/~cragin/fsQCA/download/fsQCAManual.pdf, checked on 5/4/2011.

Seelinger, Kim Thuy and Freccero, Julie (2013): Safe Haven - Sheltering Displaced Persons from Sexual and Gender-Based Violence. Comparative Report. Published by Human Rights Center University of California Berkeley School of Law. Berkeley.

Smith, Janel R.; Ho, Lara S.; Langston, Anne; Mankani, Neha; Shivshanker, Anjuli; Perera, Dhammika (2013): Clinical care for sexual assault survivors multimedia training: a mixed-methods study of effect on healthcare providers' attitudes, knowledge, confidence, and practice in humanitarian settings. In: *Confl Health* 7 (1), S. 14. DOI: 10.1186/1752-1505-7-14.

Song, Yeseul Christeena: Committing to the Future of Bangladesh: Joint Programme to address Violence against women. Key Achievements and Lessons Learned during the Intervention Period, 2010-2013. MDG Achievement Fund.

Spencer, Liz (2003): Quality in qualitative evaluation. A framework for assessing research evidence. London: Cabinet Office, Government Chief Social Researcher's Office.

Stern, Elliot; Stame, Nicoletta; Mayne, John; Forss, Kim; Davies, Rick; Befani, Barbara (2012): Broadening the Range of Designs and Methods for Impact Evaluations. Department for International Development (Working Paper, 38).

Tiwari, Agnes (2010): Effect of an Advocacy Intervention on Mental Health in Chinese Women Survivors of Intimate Partner Violence. A Randomized Controlled Trial. In: *JAMA* 304 (5), S. 536. DOI: 10.1001/jama.2010.1052.

Turnbull, B. (1999): The mediating effect of participation efficacy on evaluation use. In *Evaluation and Program Planning* 22 (2), pp. 131–140. DOI: 10.1016/S0149-7189(99)00012-9.

UNFPA-UNICEF (2012): UNICEF-UNFPA Joint Programme on Female Genital Mutilation/ Cutting: Accelerating Change. Annual Report 2012.

UNIFEM (2008): Monitoring, Evaluation and Knowledge Management Framework. UNIFEM. New York. Available online at http://www.unifem.org/attachments/products/untf_monitoring_eval_knowledge_framework.pdf.

Valovirta, V. (2002) "Evaluation Utilization as Argumentation", *Evaluation* 8: 60-80.

Weiss, Carol (1998) "Have we learned anything new about the use of evaluations?", *American Journal of Evaluation* 19: 21-33.

Weiss, Carol, E. Murphy-Graham and S. Birkeland (2005) "An alternate route to policy influence: how evaluations affect D.A.R.E." *American Journal of Evaluation* 26: 12-30.

WHO (2001) Putting women first. Ethical and safety recommendations for research on domestic violence against women. Geneva: World Health Organisation.

Williams, Suzanne (2011): Measuring Change. Synthesis of Results and Lessons from the Regional Assessment of the We Can Campaign. Published by We Can Campaign and Thoughtshop Foundation

Annex V: Interview respondents

Name of interviewee	*Role*	*Evaluated intervention*	*Country or region*	*Year (report)*
Ahmed, Julia	Implementer	Cost of Violence against Women Project	Bangladesh	2012
Askew, Ian	Commissioner	The Refentse Model for Post-Rape Care	Kenya	2009
Baluku, Moses	Implementer	We Can Campaign to End All Violence against Women	Uganda	2012
Bigirwa, Joselyn*	Commissioner and Implementer	We Can Campaign to End All Violence against Women and Prevention of Domestic Violence Project	Uganda	2012
Cappa, Claudia	Commissoner	Long-term evaluation of the Tostan Programme in Senegal	Senegal	2008
Germann, Dorsi	Evaluator	Combating Domestic Violence	West Bank	2010
Gill, Sonia	Commissioner	Strengthening Community Safety through Local Government Capacity Building	Jamaica	2011
Gunasena, Ashika	Implementer	CARE International Sri Lanka's VAW Intervention in Batticaloa: 2003-2011	Sri Lanka	2011
Hailu, Yewubdar	Evaluator	Securing the Future of Afar Pastoralist Women through Ending FGM	Ethiopia	2010
Kim, Julia	Evaluator	The Refentse Model for Post-Rape Care	Kenya	2009
Kisiero, Wilson	Implementer	CARE Refugee Assistance Project- Dabaab, Gender Based Violence Program	Kenya	2012
Moen, Hanne Lotte*	Evaluator	FOKUS FGM projects in East Africa	Ethiopia, Kenya, Tanzania	2011
Mullick, Saiqa	Commissioner	The Refentse Model for Post-Rape Care	Kenya	
Mwangi, Gladys	Evaluator	CARE Refugee Assistance Project- Dabaab, Gender Based Violence Program	Kenya	2012
Ngoma, Wendy	Commissioner	Access to Justice for Refugee Women and Girls	Tanzania	2011
Popic, Anton*	Commissioner	FOKUS FGM projects in East Africa	Ethiopia, Kenya, Tanzania	2011
Robinson, Victor	Evaluator	Cost of Violence against Women Project and CARE International Sri Lanka's VAW Intervention in Batticaloa: 2003-2011	Bangladesh and Sri Lanka	2011 and 2012
Sabbagh, Maha	Implementer	Combating Domestic Violence	Palestine	2010
Solon Helal, Isabelle	Implementer	Strategic Review of the Coalition for Women's Human Rights in Conflict	Global	2011
Thorsdalen, Sissel	Commissioner	Masimanyane Women's Support Centre	South Africa	2010
Uprety, Aruna	Evaluator	Chhaupadi Elimination Project in Achham	Nepal	2010

*Note: interviewees with an asterisk took 2 interviews, in survey preparation and Process Tracing respectively.

Annex VI: Coding instructions and reporting sheet (first round)

These coding instructions were used for the first coding round. A different set of instructions is being prepared for the second coding round.

The format has been adapted for this inception report; original reporting sheets are in 'landscape' format, which has allowed for more comfortable handling of the reporting sheets than the 'portrait' format of this report.

Preliminaries

1. Before starting with the analysis, please save the reporting sheet under the following name: "REPORTING SHEET – [name of file, without file extension & parentheses].docx"
2. If not stated otherwise, summaries of information below yes-no questions should *not* exceed approx. 50 words.
3. The answering options "almost all", "most", "some", and "almost none" correspond roughly to the following percentages: 100-76, 75 to 51, 50 to 26, 25 to 0.
4. Terms with specific definitions are underlined and italicized (for the definitions see below).
5. We recommend you start by reading all the definitions and then the questionnaire carefully at least twice, so that you have an idea of what we are asking for while you are reading the reports.
6. When working on an individual report, we recommend you go through the report once and mark all those pieces of information that might be of use. After that, check the "some orientation" section below and answer the individual questions.

Some orientation

1. For some questions vital information will typically – but not always – be found in the following parts of the report:

Part of the report	**Questions**
Title page	10a, 12
Acknowledgements	10a
Table of contents	1, 2, 2a
Executive summary	18, 33 – 38
Background	11 – 16, 19 – 24,
Methodology section	17, 25 – 32, 55 – 59, 61, 63
Findings and conclusions	39 – 42, 49
Recommendations	43, 44
Annex: Terms of Reference	11, 12, 13, 21, 23, 24
Annex: Copies of interview guidelines	50 – 54, 60

2. Answers to the following questions will most likely only be possible after you have read the whole report (either because we assume information on them is scattered throughout the report or because they require you to make judgments on the report as a whole):

Questions
3 – 7, 8 – 10, 45 – 48, 62, 64

3. The questions I – XIII should be answered in two steps: First, by making notes in the field under the respective questions and second, after you have answered all the other question of the respective section and read through your answers once again.

Definitions

- **Active evaluation stakeholders:** All those influencing the planning of the evaluation, e.g. commissioners, implementers of evaluated activities and **evaluators (!)**.
- **Active intervention stakeholders:** All those influencing the planning of evaluated activities, e.g. donors and implementers.
- **Conclusions:** Inferences drawn from findings by evaluators (interpretations, generalisations).
- **Data analysis activities:** Everything that is done to process raw data. Processing raw data involves e.g. coding, summarising and interpreting. Data analysis activities are e.g. producing a literature review, coding statistical data, qualitatively coding interviews, summarising information taken from documents.
- **Data collection activities:** Everything that is done to collect raw data (observations), partially by employing pre-designed tools such as interviews, focus group discussions, workshops, surveys, literature searches.
- **Data collection tools:** Everything that is designed for the collection of raw data (observations) such as lists of key words for online searches, interview guidelines, workshop designs, focus group discussion guidelines, standardised questionnaires.
- **Desired results:** Intended changes in the behaviour of direct and indirect beneficiaries, often called ***objectives*** or ***goals*** of the evaluated activities. Objectives are that part of the ***outcomes*** that was intended by the evaluated activities; goals are that part of the ***impact*** that was intended by the evaluated activities.
- **Direct beneficiaries:** All those receiving services (incl. training), goods or financial means as part of the evaluated activities.
- **Evaluated activities:** All those activities undertaken by implementers as part of an intervention, project or programme that the evaluation sets out to evaluate, regardless of whether they are fully funded and/or exclusively implemented by the commissioners of the evaluation (e.g. provision of services to survivors of VAWG, lobbying efforts for legislation on VAWG). All those activities that, depending on the evaluation questions/tasks, should be evaluated – not (only) activities that have been evaluated.
- **Findings:** Empirical statements. (Summary or exemplary) presentations of raw data such as quotes, facts and figures.
- **Indirect beneficiaries:** All those who experience influences of actions and products direct beneficiaries generate as part of the evaluated activities.
- **Informed non-stakeholders:** The part of the intended readership of an evaluation report which is the least involved with the implementation of evaluated activities or the evaluation. Assumed to be familiar with forms of data presentation employed by quality newspapers as well as common terminology used in planning and monitoring of development programmes. Not assumed to be familiar with social science research methodology, the specifics of the gender and VAWG field, local society and politics, or the evaluated activities.
- **Non-stakeholders:** Even though not stakeholders, for the sake of brevity referred to as a type of stakeholders in the guidelines. May be in interaction with direct beneficiaries or under positive or negative influences of actions and products of direct beneficiaries. However, those interactions/products must not be directly influenced by evaluated activities. Examples (depending on desired results of evaluated activities): Country experts, thematic experts, NGOs implementing activities similar to the evaluated activities, community leaders, members of a community in which evaluated activities were not implemented.

- **Passive intervention stakeholders:** Direct and indirect beneficiaries of the evaluated activities.
- **Qualitative evaluation questions:** Questions regarding e.g. (the existence of) types of clients, processes, the quality of cooperation.
- **Quantitative evaluation questions:** Questions regarding e.g. frequencies of interactions, numbers of beneficiaries or distributions of types of cases.
- **Recommendations:** Courses of action suggested by evaluators.
- **Terms of Reference (ToR):** When mentioned on the following pages, the terms of reference always refer to the ToR for evaluators. They usually describe the background and the purpose of the evaluation.
- **Theory of Change (ToC):** Assumed causal process linking *inter alia* evaluated activities (including tangible products or services produced) with desired results.

Questions on Evaluation Layout & Structure

NO	Question	Yes	No
1	Do the annexes include the original *Terms of Reference (ToR)*?		
2	Does the evaluation report include an executive summary?		
2a	Does the table of contents list parts of the report that are not contained in the file you are working on? If so, which parts are missing (be as specific as possible):		

NO	Question	Yes	Rather yes	Rather not	No	Not sure
3	Does the structure of the presentation of *findings* & *conclusions* in the evaluation report reflect the evaluation questions, the organisational structure of the *evaluated activities* or any other such system of categorisation that makes it easy for a reader to link the evidence to the purposes of the evaluation?					
4	Is the vocabulary employed throughout the report understandable to *informed non-stakeholders*?					
5	Does the layout allow skimming through the report (e.g. by setting key words in bold print, highlighting important conclusions, providing info boxes etc.)?					
6	Are the forms in which data is presented understandable to *informed non-stakeholders* or respective explanations given?					

NO	**Explain your answers in a short summary (ca. 250 words). (Make sure to motivate all those answers that were not a "yes" or a "no". If room permits, then go on to motivate those answers that were a very clear "yes" or "no" to you.)**
I	

NO	Question	Yes	No
II	Are there any additional aspects of layout and structure that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

Questions on the Evaluation Team

NO	Question	Yes	No
8	Is there information on past experiences of the evaluators in conducting evaluations? If yes, summarise information below:		
9	Is there information on the evaluators' experience in a similar cultural context? If yes, summarise information:		
10	Is there information on the evaluators' experience with activities in a similar field of intervention? If yes, summarise information:		

NO	Question	Yes	No
III	Are there any additional characteristics of the evaluation team that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

NO	**If the report contains e-mail addresses of the evaluator(s), the evaluation firm and/or the commissioning organisation, please copy them here.**
10a	

Questions on the Framework of the Evaluation

List evaluation questions/tasks (if there are evaluation questions on different levels of abstractness, list those on the most concrete level).
A: Indicate whether it is a *qualitative evaluation question* or a *quantitative evaluation question* or both.
B: Did the evaluation gather data that provides evidence for the question? (Three options: With several data collection tools/with a single data collection tool/no.)
PLEASE ADD A ROW TO THE TABLE FOR EACH EVALUATION QUESTION/TASK.

NO	Evaluation question/task	A: Type	B: Evidence?
11			

Is it an evaluation that was…

NO	Option
12	A: implemented by regular staff of those that implemented the *evaluated activities* (i.e. evaluators = implementers)? ☐ B: implemented by regular staff of those that financed the *evaluated activities* (i.e. evaluators = donors)? ☐ C: implemented by external consultants hired by those that **implemented** the *evaluated activities* (i.e. external evaluation)? ☐ D: implemented by external consultants hired by those that **financed** the *evaluated activities* (i.e. external evaluation)? ☐

Was the evaluation carried out…

NO	Option
13	A: at a time when at least one third of the implementation period for *evaluated activities* remained? ☐ B: at a time when less than a third of the implementation period for *evaluated activities* remained? ☐ C: at a time when the implementation period for *evaluated activities* was over? ☐ D: from the beginning to the end of the implementation period for *evaluated activities*? ☐ E: No info ☐

NO	Question	Yes	No
14	Were developments within one or several of the *active evaluation stakeholder* groups mentioned that negatively influenced the implementation of the evaluation? If yes, summarise information below:		
15	Was the cooperation of *active evaluation stakeholder* groups mentioned as a positive and/or negative influence on the implementation of the evaluation? If yes, summarise information:		
16	Is there any mention of political/social developments that negatively influenced the implementation of the evaluation? If so, summarise information below:		

NO	Question	Yes	No
17	Was the amount of resources available to evaluators mentioned as a constraint? If yes, summarise information below:		
18	Does the report include information on how findings, conclusions and/or recommendations have been disseminated or a plan as to how to do that in the future? If yes, summarise information below:		

NO	Question	Yes	No
IV	With regard to the aspects of the framework of the evaluation we have asked you to provide information on in this section, do you feel that the answering options you chose adequately reflect these aspects? If not, please qualify your answers:		

NO	Question	Yes	No
V	Are there any additional aspects of the framework of the evaluation that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

Questions on the Evaluated Activities

Describe each type of *evaluated activity* as concretely as possible. (Evaluated activities are the activities undertaken by implementers as part of an intervention, project or programme that is at the centre of an evaluation. Usually, all activities of a given intervention, project or programme are evaluated activities. Only if

evaluation questions/tasks in their entirety do clearly not cover certain activities undertaken as part of an intervention, project, or programme, are these activities not evaluated activities. Examples: Workshops/trainings/meetings, distribution of information, provision of services etc.)
A: Was the *evaluated activity* implemented in connection with other activities by the same implementer?
B: Over which period (in years) was the *evaluated activity* implemented?
PLEASE ADD A ROW TO THE TABLE FOR EACH TYPE OF EVALUATED ACTIVITY.

NO	Description of type of *evaluated activity*	A: Connected?	B: Period?
19			

NO	Where were the evaluated activities implemented? (Name country/countries. Provide number of villages/cities/regions etc., if available.)
20	

NO	Question
21	List the *desired results* (objectives and goals) that *evaluated activities* were designed to achieve.
22	Who are the main *active and passive intervention stakeholder* groups of the evaluated activities?

NO	Question	Yes	Rather yes	Rather not	No	Not sure
23	Is the *Theory of Change (ToC)* easily identifiable?					
24	Is the *ToC* (especially the link between *evaluated activities* and *desired results*) comprehensible and coherent?					

NO	Question	Yes	No
VI	With regard to the aspects of the evaluated activities we have asked you to provide information on in this section, do you feel that the answering options you chose adequately reflect these aspects? If not, please qualify your answers:		

NO	Question	Yes	No
VII	Are there any additional aspects of the evaluated activities that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

Questions on Evaluation Methodology

List the *data collection activities*. Include number of interviews/returned questionnaires/workshops, if applicable.
Insert "yes" or "no" in columns A-C in response to the following questions:
A: Is the respective sampling strategy for the implementation of the activity described and justified?
B: Are copies of guidelines/questionnaires of the respective data collection tool provided in the report or in the annex?
C: Are lists of participants/interviewees provided in the report or in the annex?
PLEASE ADD A ROW TO THE TABLE FOR EACH DATA COLLECTION ACTIVITY.

NO	*Data collection activity*	A	B	C
25				

NO	Question	Yes	No
26	Is there information on the origin/development of individual *data collection tools* (e.g. did evaluators employ standard tools, were tools tested etc.)? If so, summarise information below:		
27	Are previous research/evaluations on (some of) the same *evaluated activities* mentioned?		
28	Have any departures from the original assignment (see *ToR*, if available) been explained and justified?		
29	Have any difficulties and limitations of the methodology used and the data collected been described (including biases in data collection)?		
30	Does the report mention specific groups or individuals that were tasked with evaluation quality control (e.g reference group or review panel)?		

NO	**Is there information on how *active and passive intervention stakeholders* were involved in the design and implementation of the evaluation? If so, summarise information below:**	**Yes**	**No**
31			

NO	**Summarise additional general information on the methodology, e.g. name of overall methodological approach and justification of the methodology.**
32	

NO	Question	Yes	No
VIII	With regard to the aspects of the methodology we have asked you to provide information on in this section, do you feel that the answering options you chose adequately reflect these aspects? If not, please qualify your answers:		

NO	Question	Yes	No
IX	Are there any additional aspects of the methodology that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

Questions on Evaluation Findings, Conclusions and Recommendations

NO	Question	Yes	Rather yes	Rather not	No	Not sure
33	When looking at the executive summary only, does it allow *informed non-stakeholders* to understand the evaluators' *findings*, *conclusions*, *recommendations* and lessons learned?					

List the main *conclusions* stated in the executive summary and based on findings of the evaluation. If there is no executive summary, list those *conclusions* that are linked directly to *recommendations*.
PLEASE ADD A ROW TO THE TABLE FOR EACH CONCLUSION.

NO	Conclusion
34	

When looking at the conclusions stated in the executive summary, how many are…

NO	Option	Almost all	Most	Some	Almost none
35	...based on findings obtained through different data collection activities (triangulation of methods)?				

When looking at the *conclusions* stated in the executive summary, how many are…

NO	Option	Almost all	Most	Some	Almost none
36	...based on data obtained from *active intervention stakeholders*?				
37	…based on data obtained from *passive intervention stakeholders*?				
38	…based on data obtained from *non-stakeholders*?				

NO	Question	Yes	No
39	Are alternative *conclusions* (i.e. contradictory *conclusions* based on the same data) presented and discussed?		

NO	Option	Almost all	Most	Some	Almost none
40	When looking at the *findings* as a whole, how many would you say are appropriately sourced (i.e. the respective literature, interview or survey is given)?				

NO	Question	Yes	No
41	Are contradictory or unexpected *findings* presented and discussed?		
42	Is the analysis of data disaggregated to show outcomes and impact of the *evaluated activities* on different *passive intervention stakeholder groups*?		
43	Do *recommendations* explicitly refer to the resources necessary to implement them (time, people, money)?		
44	Do *recommendations* specify actors responsible for their implementation?		

NO	Question	Yes	No
45	Have any stakeholders been given opportunities to comment on the *conclusions*, *recommendations* and lessons? If so, summarize information on these opportunities below:		
46	Are gender issues, as well as the ways in which other aspects of identity (such as age, class, race, religion and minority status) influence gender imbalances discussed? If so, summarise the discussion below (max. 100 words):		
47	Does the evaluation include an analysis of the interdependencies of power and gender? If so, summarise the discussion below:		
48	Are characteristics of social norms and structures in the geographical region in which the *evaluated activities* were implemented discussed? If so, summarise the discussion below (max. 100 words):		

NO	Question	Yes	No
X	With regard to the aspects of the findings, conclusions and recommendations we have asked you to provide information on in this section, do you feel that the answering options you chose adequately reflect these aspects? If not, please qualify your answers:		

NO	Question	Yes	No
XI	Are there any additional aspects of the findings, conclusions and recommendations that you feel could influence the effect of the evaluation, but that we have not asked you to provide information on? If yes, please add information on these aspects:		

Questions on Ethical Considerations

NO	Question	Yes	No
49	Are there instances in which informants (individuals or small groups) are identifiable as individuals/individual local groups? If so, summarise information below:		

NO	Question	Yes	No	No info
50	Have stakeholders who were asked to provide information been informed of the purpose of the evaluation?			
51	Have stakeholders who were asked to provide information been informed of their right to refuse participation in the data collection process?			
52	Have stakeholders who were asked to provide information been offered anonymity?			
53	Does the report use pictures of individuals?			
54	If so, does the report indicate that people have been asked permission to be photographed and fully informed as to how, where and when pictures will be shown?			

NO	Question	Yes	No	No info
55	Have there been any interviews with VAWG survivors who were identified as such?			
56	If so, were VAWG survivors identified in a way to prevent that any community member could identify them as such?			
57	If so, and if other members of the survivor's household or of the community were interviewed, did the interview include questions on attitudes to, experience with or use of VAWG?			
58	Have those collecting or analysing data on VAWG survivors been recruited from outside the survivors' social networks?			
59	If VAWG survivors were interviewed, have potential providers of support (health, legal, social services, women's organisation) been identified prior to the research and a list of reliable providers been prepared for sharing with interviewees and/or has a trained counsellor accompanied the interviewers to provide support as needed?			

NO	Question	Yes	No	No info
60	Are all questions about VAWG and its consequences asked in a supportive and non-judgmental manner, in a language that cannot be interpreted as blaming or stigmatising?			
61	Have data (interview recordings, questionnaires etc.) been stored safely?			

NO	Question	Yes	No
63	Have interviewers (including the evaluators) had specific training to conduct interviews with VAWG survivors in a way to prevent harmful effects on interviewees? If so, summarise information below:		

NO	Question	Yes	No
64	Are the vocabulary employed and the form in which data is presented respectful of those who information is provided on and free of gender bias? If not, why?		

NO	Question	Yes	No
XII	Have other cultural, ethical and legal concerns been taken into account in the design of data collection tools and in the implementation of data collection activities that are not covered by the questions above? If so, summarise the information below:		

Annex VII: Survey questions

The survey was administered as a web-based survey using LimeSurvey, a highly adaptive survey design software, and LimeService, an online survey hosting service. The list of questions as presented here appears repetitive. This is because depending on their role in the evaluation, respondents have received different sets of questions. The distribution of these questions is outlined in the introductory table ("Thematic structure of the survey").

For this report, we have chosen to present the questions in this form, which reveals the logic of the questionnaire. The layout of the web-based survey is different and very user-friendly. A PDF copy of the review in its web-based layout can be made available upon request.

Thematic structure of the survey

Topic	Items	E*	C	F	I
Welcome note	W1	0	0	0	0
Note	N1	0	0	0	0
Prelude	P1	1	1	1	1
Resources	R1 – R5	5	2	2	2
Professional standing and sensitivities	PS1 – PS6	2	5	5	5
Findings and available data	F1, F2	1	1	1	1
Participatory (design and analysis)	PD1 – PD4	4	0	0	0
Timeliness	T1 – T3	2	0	0	2
Dissemination	D1 – D4	0	5	5	5
Political environment	PE1 – PE7	2	4	4	6
Effects	E1 – E7	0	5	8	6
Thank you	G1	0	0	0	0
	Max. number of questions	17	23	26	28

*E = evaluators, C = commissioners, F = funders, I = implementers.

Question list

Code	Question	E	C	F	I

Code	Question	E	C	F	I
D1	Who did your organization distribute the final evaluation report to and how? For each group of people, choose all those distribution channels that apply. Representatives of the organisation(s) that implemented the evaluated intervention (email/hard copy/personal presentation/not distributed to/don't know) Representatives of the organisation(s) that funded the evaluated intervention (email/hard copy/personal presentation/not distributed to/don't know) Representatives of organisations that implement interventions similar to those evaluated (email/hard copy/personal presentation/not distributed to/don't know) Representatives of organisations that fund interventions similar to those evaluated (email/hard copy/personal presentation/not distributed to/don't know) Beneficiaries of the evaluated intervention (email/hard copy/personal presentation/not distributed to/don't know) Representatives of government and other public agencies regulating a field related to the evaluated intervention (email/hard copy/personal presentation/not distributed to/don't know)		X	X	X
D2	In which other ways did your organisation distribute the final evaluation report? Own websites or blogs (yes/no) Other websites or blogs (yes/no) Mailing lists (yes/no) Social media (Facebook, Twitter, LinkedIn etc.) (yes/no)		X	X	X
D3	How often did your organisation use the evaluation and its findings in the year after the evaluation was completed? The findings of the evaluation were discussed in internal documents, meetings and/or workshops. (Never/Once/Several times/Don't know) The findings of the evaluation were discussed in documents disseminated to a selected audience. (Never/Once/Several times/Don't know) The findings of the evaluation were discussed in meetings and workshops with a selected audience. (Never/Once/Several times/Don't know) The findings of the evaluation were discussed in documents disseminated to the wider public. (Never/Once/Several times/Don't know) The findings of the evaluation were discussed in public events. (Never/Once/Several times/Don't know)		X	X	X

Code	Question	E	C	F	I
D4	How often did the media (e.g. press, radio, TV) report on the evaluation or its findings in the year after the evaluation was completed? The findings of the evaluation were discussed in the local media. (Never/Once/Several times/Don't know) The findings of the evaluation were discussed in national and/or international media. (Never/Once/Several times/Don't know)		X	X	X
D5	Overall, how much did the evaluation influence organisations working on VAWG that were not directly involved in the evaluated intervention? The evaluation has influenced the way they design and/or implement evaluations. (Not at all/A little/Somewhat/Strongly) The evaluation has influenced the way they design and/or implement interventions. (Not at all/A little/Somewhat/Strongly) The evaluation has influenced their organizational structures and processes. (Not at all/A little/Somewhat/Strongly)		X	X	X
E1	How much do you agree with the following statements regarding the findings presented in the final evaluation report? The evaluation findings have confirmed our understanding of the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation findings have changed our understanding of the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) Some findings of the evaluation were surprising to us. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation report describes the evaluated intervention in an accurate manner. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The conclusions presented in the evaluation report are credible judgments on the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)	N	X	X	X

Code	Question	E	C	F	I
E2	Regarding the effect of the evaluation on the way your organisation designs and implements interventions, how much do you agree with the following statements? The evaluation helped us to change the way we implemented the evaluated intervention and/or follow-up interventions. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to change the way we implement interventions in general. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation helped us to maintain the way we implemented the evaluated intervention and/or follow-up interventions. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to maintain the way we implement interventions in general. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)				X
E3	Regarding the effect of the evaluation on the way your organisation collaborates with implementing organisations, how much do you agree with the following statements? The evaluation helped us to change the way we collaborated with the organisation that implemented the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to change the way we collaborate with implementing organisations in general. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation helped us to maintain the way we collaborated with the organisation that implemented the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to maintain the way we collaborate with implementing organisations in general. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)			X	

Code	Question	E	C	F	I
E4	Below are pairs of statements. How would you place your views on the scale, where 1 means you completely agree with the statement on the left and 10 means you completely agree with the statement on the right? If your views fall somewhere in between, you can chose a number in between. The evaluation contributed to a situation in which… …consultations with the beneficiaries of the intervention became less regular./ …consultations with the beneficiaries of the intervention became more regular. …our beneficiaries became more confused with regard to what the intervention was all about./ …our beneficiaries gained a better understanding of what the intervention was all about. …beneficiaries became more hesitant to make their voices heard./ …beneficiaries felt more entitled to have their voices heard.		X	X	X
E5	How much do you agree with the following statements on the effect of the evaluation on the persons and organisations which (at the time) supported your organisation financially? The evaluation has helped to maintain their financial support to our organisation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to increase their financial support to our organisation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has caused hesitation among them to maintain their financial support to our organisation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has led to a decrease of their support to our organisation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)		X	X	X

Code	Question	E	C	F	I
E6	Regarding your collaboration with the organisation(s) that implemented the evaluated intervention, how much do you agree with the following statements? The evaluation has helped to maintain our financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to increase our financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has caused hesitation among us to maintain our financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has led to a decrease of our support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)			X	
E7	And how about the effect of the evaluation on the collaboration of the implementing organisation(s) with other persons and organisations that (at the time) supported it/them financially? The evaluation has helped to maintain their financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to increase their financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has caused hesitation among them to maintain our financial support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has led to a decrease of their support to the organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)			X	

Code	Question	E	C	F	I
E8	Regarding the effect of the evaluation on the networks of your organisation, how much do you agree with the following statements? The evaluation has helped us to come into contact with other organisations that could fund our work. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to come into contact with other organisations that we could collaborate with locally. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped us to come into contact with public institutions. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to improve existing cooperation with organisations that have funded our work. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to improve existing cooperation with organisations that we have collaborated with locally. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluation has helped to improve existing cooperation with public institutions. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)		X	X	X
E9	Overall, how much are you satisfied with the following aspects of the evaluation? Please place your views on the scale, where 1 means you are totally dissatisfied and 10 means you are totally satisfied with the respective aspect. Evaluation design Manner in which the evaluator(s) conducted the evaluation Quality of the final evaluation report	N	X	X	X
F1	Do you remember the final evaluation report to be rather negative or rather positive about the evaluated intervention? (Very negative/Rather negative/Mixed/Rather positive/Very positive)		X	X	X
F2	Before work on the evaluation started, were there any other evaluations or research projects on the same intervention or parts thereof (e.g. baseline studies, mid-term reviews)? (Yes/No/Don't know.)	X			
G1	Thank you very much for your participation. If you have questions regarding the survey or would like to give your feedback, do not hesitate to contact us on review-team@gmx.de. You can also visit our blog [html link].	X	X	X	X

Code	Question	E	C	F	I
N1	Please note: All questions in this questionnaire relate to the following evaluation: [Token: Evaluation] "Evaluated intervention" refers to the specific project, campaign, program, initiative or a certain period of an organisation's activities the evaluation was supposed to evaluate. With "implementing organization" we will refer to the organisation that implemented the evaluated intervention.	X	X	X	X
P1	In which way(s) have you been involved in the evaluation? Please chose all options that apply. (I was the evaluator or one of the evaluators./I am/was part of an organization that commissioned the evaluation./ I am/was part of an organization that has provided funding to the project, programme, organization or initiative that has been evaluated./ I am/was part of an organization that was evaluated or that implemented (part of) the project, programme or initiative that was evaluated.)	X	X	X	X
PD1	During the evaluation, were there any meetings and/or workshops to discuss preliminary findings, conclusions and recommendations of the evaluators? (No/Yes)	X			
PD2	Who participated in those meetings and/or workshops? Beneficiaries of the evaluated intervention (No/Yes) Members of organisations whose work was evaluated (No/Yes) Members of organisations that funded the evaluated intervention (No/Yes)	X			
PD3	Were there any meetings and/or workshops to discuss the evaluation design or the data collection process? (No/Yes)	X			
PD4	Who participated in those meetings and/or workshops? Beneficiaries of the evaluated intervention (No/Yes) Members of organisations whose work was evaluated (No/Yes) Members of organisations that funded the evaluated intervention (No/Yes	X			
PE1	Regarding the collaboration with stakeholders of the evaluation, how much do you agree with the following statements? The implementing organisation(s) fully supported my work. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The funding organisation(s) fully supported my work. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)	X			

Code	Question	E	C	F	I
PE2	Regarding the collaboration with the evaluator(s), how much do you agree with the following statements? The evaluator(s) communicated with implementing organisation(s) in an appropriate and unambiguous way. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluator(s) communicated with funding organisation(s) in an appropriate and unambiguous way. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluator(s) was/were open to suggestions by staff of implementing organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The evaluator(s) was/were open to suggestions by staff of funding organisation(s). (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)	N	X	X	X
PE3	How much did aspects of the political, legal and social contexts in which the intervention took place have a negative impact on the realisation or use of the evaluation? (Not at all/A little/Somewhat/Strongly)				X
PE4	In your opinion, what were the reasons for the evaluation? The list below shows seven possible reasons. Please choose the reasons that apply by dragging them to the right and rank them according to their importance, placing the most important reason on top of the column. Evaluation requirements of the funding organisation(s) Evaluation requirements of the implementing organisation(s) Information needs of the implementing organisation(s) Information needs of the funding organisation(s) Need of implementing organisation(s) to reflect on practices Need of funding organisation(s) to reflect on practices Need to solve concrete organisational problems	X	X	X	X

Code	Question	E	C	F	I
PE5	In the period from the beginning of the evaluation until one year after the evaluation was completed, did your organisation undergo changes with regard to the following aspects? Which role, if any, did the evaluation play in this? Income (grants, donations, service fees etc.) (Major decrease/Minor decrease/No changes/Minor increase/Major increase/Don't know) (No role/minor role/major role.) Size of the workforce (Major decrease/Minor decrease/No changes/Minor increase/Major increase/Don't know) (No role/minor role/major role.) Scope of activities (Major decrease/Minor decrease/No changes/Minor increase/Major increase/Don't know) (No role/minor role/major role.)		X	X	X
PE6	In the period from the beginning of the evaluation until one year after the evaluation was completed, did your organization undergo changes with regard to the following aspects? Which role, if any, did the evaluation play in this? Leadership of the organization (No changes/Minor changes/Major changes) (No role/minor role/major role.) Themes the organization worked on (No changes/Minor changes/Major changes) (No role/minor role/major role.)		X	X	X
PE7	When you think of the time shortly before the evaluators began their work, how much do you agree with the following statements? Our organisation had a clear understanding of what we wanted to learn from the evaluation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The members of our organisation who implemented the evaluated intervention welcomed the prospect of having an evaluation. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.) The decision to conduct an evaluation caused uneasiness among the members of our organisation who implemented the evaluated intervention. (Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree/Don't know.)				X
PS1	How many evaluations and similar research assignments had the evaluator/the most experienced member of the evaluation team carried out before this evaluation? If you are not sure, please choose the option that seems most likely. (None/Up to 5/6 to 20/More than 20)	X	X	X	
PS2	In your opinion, how familiar was the evaluator/evaluation team with research methods? (Don't know/Not familiar at all/Rather unfamiliar/Rather familiar/Very familiar)	N	X	X	X
PS3	In your opinion, how familiar was the evaluator/evaluation team with the local political, social and cultural context in which the evaluated intervention took place? (Don't know/Not familiar at all/Rather unfamiliar/Rather familiar/Very familiar)	N	X	X	X

Code	Question	E	C	F	I
PS4	In your opinion, how familiar was the evaluator/evaluation team with the type of intervention/organisation that was evaluated? (Don't know/Not familiar at all/Rather unfamiliar/Rather familiar/Very familiar)	N	X	X	X
PS5	In your opinion, how familiar was the evaluator/evaluation team with gender research? (Don't know/Not familiar at all/Rather unfamiliar/Rather familiar/Very familiar)	N	X	X	X
PS6	It can be risky to undertake evaluations in the field of violence against women and girls (VAWG). How likely were the following risks with regard to the intervention that was been evaluated? Beneficiaries, activists and people in their social environment could be threatened, brutalised or marginalised if others learned that they talked about their experience with VAWG or VAWG-related work. (Very unlikely/Unlikely/Likely/Very likely/Don't know.) Beneficiaries who survived VAWG could suffer psychological harm if they were questioned personally about anything related to their experience of VAWG. (Very unlikely/Unlikely/Likely/Very likely/Don't know.) Researchers could experience harmful distress when undertaking an evaluation of the intervention. (Very unlikely/Unlikely/Likely/Very likely/Don't know.)	X			X
R1	What was the approximate total budget for the evaluation (in US Dollars)?? If you are not sure, please choose the option that seems most likely. (No extra resources allocated to the evaluation/Up to 5000$/Between 5001$ and 10000$/Between 10001$ and 50000$/More than 50000$)	X	X	X	X
R2	In your opinion, was the budget sufficient given the extent of the evaluation tasks? Please use the scale below, where 1 means that you completely agree with the first statement and 10 means you completely agree with the second statement. If your views fall somewhere in between, you can chose any number in between. The budget was too small to do everything that would have been necessary to fulfill the evaluation tasks. The budget allowed to do a lot more than would have been necessary to fulfill the evaluation tasks.	X	X	X	X
R3	How many persons received honoraria for the implementation of the evaluation? (Don't know)	X			
R4	How many persons carried out activities to gather or process information for the evaluation (e.g. by conducting interviews or focus group discussions or by imputing data)? (Don't know)	X			
R5	How much time (in full months) passed between the moment work on the evaluation commenced, and the delivery of the final report? If you are not sure, please estimate. (Don't know)	X			

Code	Question	E	C	F	I
T1	In your opinion, would the evaluation have benefitted from taking place earlier or later than it actually did? (No/Yes, if earlier./Yes, if later./Don't know.)	X			X
T2	Why would the evaluation have benefitted from taking place earlier?	X			X
T3	Why would the evaluation have benefitted from taking place later?	X			X
W1	Welcome! You are invited to answer questions about an evaluation that is part of our review of evaluations related to violence against women and girls. The review has been commissioned by the UK Department for International Development (DFID); it analyses 74 evaluations. The purpose is to learn about evaluation methods and the way evaluations have been used. All data will be presented in a way that makes it impossible to trace our findings to any individual evaluations or specific respondents. Navigation: Please use the buttons "next" and "previous" on the page to navigate through the questionnaire. Please do not use the back and forward buttons of your browser.	X	X	X	X

Annex VIII: Coding instructions (second round)

For the second coding round, each coder was assigned one of three work packages. Coder-specific information has been replaced by "[coder-specific]".

Work package 1:

Overview

Two types of tasks have been prepared for you:

1. Re-coding specific aspects for [coder-specific] reports you have worked on in the 1st coding round
2. Coding some aspects of [coder-specific] reports you have not worked on before

We have foreseen a total of [coder-specific] full working hours for the completion of these tasks.

You have been invited to a dropbox folder which includes the following items:

- [coder-specific] evaluation reports you have worked on
- [coder-specific] evaluation reports that are new to you
- A separate folder with TOR or similar documents for evaluation reports that do not include the TOR in the actual report
- EXCEL tables that you will need for coding
- A sub-folder where to store your work

Please call Wolf or send us an e-mail message (Wolf cc. Michaela) whenever any question arises. You are invited to **store all work-in-progress for this assignment in our shared dropbox folde**r so that we can monitor it and get in touch with you in case anything needs to be corrected.

A. Recoding 'your' 1st round reports

The following three tasks are to be performed on the following reports only: [coder-specific]. Please record all data on the EXCEL sheets provided in 'your' dropbox folder.

Coding evaluation reports: data used in conclusions

1. For each of the evaluation reports listed above, study the conclusions stated in the executive summary. Please assess: When looking at the conclusions stated in the executive summary, how many are…

- based on ***original data*** (collected as part of the evaluation with data collection tools provided by the evaluator(s)).
- based on ***original analysis*** (data collected by others with data collection tools provided by others, analysed by the evaluator(s)).
- based on ***documents*** (data analysis cited by the evaluator(s)).

2. Code your assessment in the respective columns of EXCEL sheet C – [coder-specific] ("original data", "original analysis" and "documents") as follows:

- Almost all conclusions = 4
- Most conclusions = 3
- Some conclusions = 2
- Almost none = 1

Save the EXCEL file under its original name.

Coding 1st round reporting sheets: types of data

1. Create a **WORD document** for each evaluation report in **"EXCEL sheet C –[coder-specific]"**. Name it in the following way: "[Last name of main author] [year of publication] – method – [your first name]". Take the information for last name of main author and year of publication from the EXCEL sheet C. For instance, "Chibuta 2011 – method – Scout.doc".

2. In each WORD document, create a table with two columns. Name the columns "Data collection tool" and "Type of data" respectively.

3. For each of the [coder-specific] evaluations, study **question 25 in the relevant 1st coding round reporting sheet**. Please determine, for every data collection tool, whether the tool has provided preponderantly closed-ended data or open-ended data. Create a new line in the document for each data collection tool.

KEY DEFINITIONS

Closed-ended data: Interviewer frames question and answer options. Closed-ended data is typically collected to confirm (or discard) hypotheses. Closed-ended data is usually generated through such tools as surveys, check lists, structured interviews, and safety audits.

Open-ended data: Interviewer frames topics/aspects of topics or key questions, but leaves open exact wording and answer options. Open-ended data is typically collected to explore/generate new hypotheses. Open-ended data tends to be generated in conversations, focus group discussions, other group discussions, open interviews, biographic interviews.

For tools not listed above, check the evaluation report to assess whether they yield preponderantly closed-ended or open-ended data. If no information = unclear.

- If data collection tool produced predominantly closed-ended data, write "1" in second column.
- If data collection tool produced predominantly open-ended data, write "0" in second column.
- If unclear, write "9".

Save each completed WORD document under its original name.

3. Use **EXCEL sheet C – [coder-specific], in particular the columns** "open-ended" and "closed-ended".

4. Study the **conclusions in the executive summary of the actual evaluation reports**. Take into account the importance each type of data collection tool has had for the conclusions, and whether the tools have produced preponderantly open-ended or closed-ended data. How many conclusions are *based on open-ended data*, *based on closed-ended* data?

Write the results the respective columns, coding them as follows:

- Almost all conclusions= 4
- Most = 3
- Some = 2
- Almost none = 1

Save the EXCEL file under its original name.

Coding TOR/ evaluation reports: beneficiaries, objectives, theory of change

1. Use **EXCEL sheet C – [coder-specific]**, in particular the columns "def. of beneficiaries", "objectives", and "TOC".

2. Check the evaluation Terms of Reference (TOR) to answer the question: Is the population that is supposed to benefit from the intervention well defined?

- ***Well defined* (score = 2)** includes information on sex, age group, geographical region and number of beneficiaries.
- ***Adequately defined* (score = 1)** includes information on sex and number of beneficiaries.
- ***Poorly/ not defined* (score = 0)** includes information on number or sex only, or no information at all.

Write the respective scores in the column "def. of beneficiaries".

3. Check the TOR to answer the question: Are the objectives or expected outcomes of the intervention stated in clear terms?

The objectives or expected outcomes are:

***Well defined* (score = 2)** if they fully meet all of the following characteristics: specific, measurable, and attainable.

- **Specific**: They state in reasonably precise terms what change the intervention is supposed to accomplish or contribute to *(for instance, 'improved access to legal aid' – not 'a life free of violence')*
- **Measurable**: They come with indicators for qualitative or quantitative measurement of progress
- **Attainable**: they can realistically be achieved with the resources and time-frame of the project *(not 'a life free of violence')*

***Vaguely defined* (score = 1)** if they have 2 of the characteristics listed above.

***Poorly defined* (score = 0)** if they have up to 1 of the characteristics.

Write the respective numbers in the column "objectives".

4. Check the TOR to answer the question: Is the way in which the intended objectives or outcomes will be achieved (theory of change, TOC) described in a coherent manner?

- ***Coherent*** if cause-to-effect links are clearly described and ***factors external to the intervention*** (context) are taken into account (score = 2).
- ***Deficient*** if cause-to-effect links are described ***without taking into account context factors*** (score = 1).
- ***Absent*** if there is ***no explanation*** of the links between activities and intended outcomes/ objectives (score = 0).

Write the respective scores in the column "TOC" and save the EXCEL file under its original name.

B. Coding a 'new' set of reports

The following tasks are to be performed on the following reports and the corresponding reporting sheets: [coder-specific]. Please record all data on the EXCEL sheets provided in 'your' dropbox folder.

Assessing presentation

1. Use **EXCEL sheet B – [coder-specific]** and open the empty reporting sheet (in your dropbox folder) to look up the corresponding questions. **Answer questions 2, 3, 4, 5, 6, 33, 43 and 44** of the reporting sheet for each evaluation report in EXCEL sheet B.

Write the answers into the respective columns of the EXCEL table. Write down only numbers, as follows:

- If answer options yes/no: yes = 1, no = 0.
- If answer options yes/rather yes etc.:
    - yes = 4
    - rather yes = 3
    - rather no = 2
    - no = 1
    - not sure = 9
    - no answer = 99

2. **Check List A** for additional tasks. (Please note that List A will be added to your folder at a later point.) Create columns as described in List A and answer the respective questions, again using only numbers, as described in List A.

Bias assessment

1. **Create a WORD document for each evaluation report** listed in EXCEL sheet B - Scout. Name the document as follows: "[Last name of main author] [year of publication] – design – [your first name]". Take the information for last name of main author and year of publication from EXCEL sheet B - Scout. Example: "*Moen 2012 – design – Scout.doc"*.

2. In each WORD document, **create a table with seven columns**. Name the columns as follows, from left to right: "Group of informants", "Data collection method", "Selection bias", "Justification selection bias", "Power bias", "Justification power bias" and "Type of stakeholder".

3. All informants the evaluators have asked with the same data collection tool are one group. Different data collection tool => different group. (If, for instance, the whole evaluation is based on a single data collection tool, all informants form just one group.)

   Check #25 of the respective reporting sheets for information on data collection tools.

   If you are not sure whether the information in #25 is complete or correct, check the evaluation report, and add or regroup informants accordingly. Name the groups of informants in such a way that differences between them are recognisable. Example: "Staff of implementing organisation, and justice and police personnel who have participated in training courses".

   Start a new line in the first column of the table for each group of informants and state the data collection tool in the second column.

KEY DEFINITIONS

Selection bias is defined as follows: We define **selection bias** as problems related to the objectivity and fairness of data collection which are caused by the way in which people who provide information (informants) are chosen.

Selection bias is introduced if (1a) only active intervention stakeholders are informants, and/or (1b) informants are selected exclusively by active intervention stakeholders. Such bias can be prevented if substantial information is gathered from informants who are (2a) randomly selected or (2b) selected by the evaluation team according to a scheme which factors in their different relationships to the evaluated intervention.

Power bias is defined as follows: We define **power bias** as problems related to the objectivity and fairness of data collection which are caused by placing people who provide information (informants) into situations where unequal power relationships between those present make it difficult for some or all participants to express themselves freely.

For instance, in societies where power dynamics expect women to remain silent or to avoid publicly contradicting men, focus groups bringing together women and men would induce power bias. Similar risks exist when beneficiaries are interviewed by staff of the implementing organisation, or in the presence of such staff. Power dynamics bias is prevented if (4a) respondents may provide information anonymously.

3. Read **definition of selection bias**. Provide an assessment for each group of informants **whether selection bias was**:

    a. absent/very weak (0)
    b. weak/rather weak (1)
    c. rather strong/strong (2)
    d. very strong (3)

    Insert the respective number into the third column.

    Justify your assessment in max. 50 words for each group of informants, in the next column to the right.

4. Read **definition of power bias**. Provide an assessment for each group of informants **whether power bias was**:

    a. absent/very weak (0)
    b. weak/rather weak (1)
    c. rather strong/strong (2)
    d. very strong (3)

    Insert the respective number into the third column.

    Justify your assessment in up to 50 words for each group of informants, in the next column to the right.

5. There are three types of stakeholders: **active intervention stakeholders, passive intervention stakeholders, non-stakeholders. Check the reporting sheet for definitions.** For each group of informants you have listed, determine which type of stakeholder they represent and write the corresponding type of stakeholder into column 7.

Assessing elements that contribute to the complexity of the evaluation task

1. Use **EXCEL sheet B** - [coder-specific].

2. Check **#11 of respective reporting sheet**. The evaluation questions/tasks indicate **the object of the evaluation –** i.e. what the evaluation set out to evaluate. This could encompass a wide range of aspects, such as all activities an organisation is involved in, or something much more limited, such as an organisation's policies and procedures. **If unclear, check in the evaluation report**.

   Assess **whether the object of evaluation includes direct work with private individuals,** for example, the provision of services for VAWG survivors. **If yes, write "1" in column "beneficiaries". If no, write "0" in column "beneficiaries".** (For example: the policies of an organisation that works with VAWG survivors.)

3. **Check whether assessing "impact" is mentioned among the evaluation tasks**. This is the case if "impact" is explicitly mentioned as something that must be assessed. **If yes, write "1" in column "impact". If no, write "0" in column "impact".**

Assessing the basis of evaluation findings

1. Use EXCEL sheet B - [coder-specific].

2. **Answer questions 35, 36, 37, 38 for each evaluation report** in excel file B. Write down only numbers:

   - Almost all = 4
   - Most = 3
   - Some = 2
   - Almost none = 1.

3. **When looking at the conclusions stated in the executive summary, how many are *based***

- *on **original data*** (collected with data collection tools designed by the evaluator(s))
- *on **original analysis*** (data collected with data collection tools designed by others, analysed by the evaluator(s)
- *on **documents*** (data analysis cited by the evaluator(s)

Write **numbers** in the respective columns ("original data", "original analysis" and "documents"):

- Almost all conclusions = 4
- most = 3
- some = 2
- almost none = 1

Assessing types of data collected

1. **Create a WORD document** for each evaluation report in **Excel file B**. Name it in the following way: "[Last name of main author] [year of publication] – method – [your first name]". Take the information for last name of main author and year of publication from Excel file B. Example: "Moen 2012 – method – Scout.doc".

2. In each WORD document, **create a table with two columns**. Name them "Data collection tool" and "Type of data".

3. **Check question 25 in the reporting sheets** to determine, for every data collection tool, whether it provided preponderantly closed-ended data or open-ended data (as defined above, p.2). **For tools not listed in the definitions, check the evaluation report** to assess whether they yield preponderantly closed-ended or open-ended data. If no info = unclear.

    - If data collection tool produced **predominantly closed-ended data, write "1"** in second column.
    - If data collection tool **produced predominantly open-ended data, write "0"** in second column.
    - If **unclear, write "9".**

4. Use **EXCEL sheet B –** [coder-specific], in particular the columns "open-ended" and "closed-ended".

5. Study the **conclusions in the executive summary of the actual evaluation reports**. Take into account the importance each type of data collection tool has had for the conclusions, and whether the tools have produced preponderantly open-ended or closed-ended data. How many conclusions are *based on open-ended data*, *based on closed-ended* data? Write numbers in the respective columns:

    - Almost all conclusions = 4
    - most = 3
    - some = 2
    - almost none = 1

Assessing aspects of project design

For this task, you will work chiefly on the TOR (annexed to the report or in a separate file).

1. **Use EXCEL file B –** [coder-specific], in particular the columns "def of beneficiaries", "objectives", "TOC".

2. **Check the TOR to answer the question: Is the population** that is supposed to benefit from the intervention well defined?

    - ***Well defined*** (score = 2) includes information on sex, age group, geographical region and number of beneficiaries.
    - ***Adequately defined*** (score = 1) includes information on sex and number of beneficiaries.
    - ***Poorly/ not defined*** (score = 0) includes information on number or sex only, or no information at all.

Write the respective scores in column “def of benecificaries”.

3. **Check the ToR to answer the question: Are the objectives or expected outcomes** of the intervention stated in clear terms?
   - ***Well defined*** **(score = 2)** if they fully meet all of the following characteristics: specific, measurable, and attainable.
     - **Specific**: They state in reasonably precise terms what change the intervention is supposed to accomplish or contribute to *(for instance, ‘improved access to legal aid’ – not ‘a life free of violence’)*
     - **Measurable**: They come with indicators for qualitative or quantitative measurement of progress
     - **Attainable**: they can realistically be achieved with the resources and time-frame of the project *(not ‘a life free of violence’)*
   - ***Vaguely defined*** **(score = 1)** if they have 2 of the characteristics listed above.
   - ***Poorly defined*** **(score = 0)** if they have up to 1 of the characteristics.

   Write the respective scores in the column “objectives”.

4. **Check the TOR** to answer the question: Is the way in which the intended objectives or outcomes will be achieved (**theory of change**, TOC) described in a coherent manner?
   - ***Coherent i***f cause-to-effect links are clearly described, and ***factors external to the intervention*** (context) are taken into account (score = 2).
   - ***Deficient*** if cause-to-effect links are described ***without taking into account context factors*** (score = 1).
   - ***Absent*** if there is ***no explanation*** of the links between activities and intended outcomes/ objectives (score = 0).

   Write the respective scores in the column “ToC”.

Work package 2:

Overview

Four types of tasks have been prepared for you:

3. Working with data on “presentation” from the 1st coding round reporting sheets for the 39 evaluations in our QCA set. **(Please begin with this 1st task as soon as possible and complete it before you move to the following tasks.)**
4. Re-coding specific aspects for [coder-specific] reports you have worked on in the 1st coding round
5. Google search on the publication status of the 39 evaluations
6. Coding aspects of other coders’ 2nd round reports. (That task will have to come after all other coders will have completed their coding work.)

We have foreseen a total of [coder-specific] full working hours for the completion of all tasks described above.

You have been invited to a dropbox folder which includes the following items:

- The 39 evaluation reports in our QCA set

- The 1st coding round reporting sheets for those 39 reports
- A separate folder with TOR or similar documents for evaluation reports that do not include the TOR in the actual report
- EXCEL tables that you will need for coding
- A sub-folder where to store your work

Please call Wolf or send us an e-mail message (Wolf cc. Michaela) whenever any question arises. You are invited to **store all work-in-progress for this assignment in our shared dropbox folde**r so that we can monitor it and get in touch with you in case anything needs to be corrected.

A. Priority task: Coding 1st round data on presentation and measurement

This task should begin as soon as possible and be completed before you start with any other tasks.

1. Read items "I", "II" and "X" of all 39 reporting sheets. Use **EXCEL Sheet A -** [coder-specific] (EXCEL sheet provided to you) to note down all issues with existing measurements and proposals for additional measurements that are mentioned for each evaluation. For the responses pertaining to "X", write down only those relating to tasks 43 and 44 in the reporting sheet. Save the completed list under the original name.

2. Based on the data gathered in the first step above, reflect on all issues that you have found. Please write up a summary of (a) the issues with existing measurements you have found in the reporting sheets and (b) proposals for additional measurements. With regard to (a), make your own proposals as to how the respective measurements could be improved. The summary and your proposals should not exceed 400 words. Please save the text under the name **MEASUREMENT –** [coder-specific]**.doc** and send an e-mail message to Wolf and Michaela to let us know the task is completed

B. Recoding 'your' 1st round reports

The following three tasks are to be performed on the following reports only: [coder-specific]. Please record all data on the EXCEL sheets provided in 'your' dropbox folder.

Coding evaluation reports: data used in conclusions

1. For each of the evaluation reports listed above, study the **conclusions stated in the executive summary**. Please assess: When looking at the conclusions stated in the executive summary, **how many are…**

- based on ***original data*** (collected as part of the evaluation with data collection tools provided by the evaluator(s)).
- based on ***original analysis*** (data collected by others with data collection tools provided by others, analysed by the evaluator(s)).
- based on ***documents*** (data analysis cited by the evaluator(s)).

2. Code your assessment in the respective columns of EXCEL sheet C – [coder-specific] ("original data", "original analysis" and "documents") as follows:

- Almost all conclusions = 4
- Most conclusions = 3

- Some conclusions = 2
- Almost none = 1

Save the EXCEL file under its original name.

Coding 1st round reporting sheets: types of data

1. Create a **WORD document** for each evaluation report in **"EXCEL sheet C –**[coder-specific]**"**. Name it in the following way: "[Last name of main author] [year of publication] – method – [your first name]". Take the information for last name of main author and year of publication from the EXCEL sheet C. For instance, "Chibuta 2011 – method – Paula.doc".
2. In each WORD document, create a table with **two columns. Name the columns "Data collection tool" and "Type of data"** respectively.
3. For each of the 5 evaluations, study **question 25 in the 1st coding round reporting sheets for the respective evaluation**. Please determine, for every data collection tool, whether the tool has provided preponderantly closed-ended data or open-ended data. **Create a new line in the document for each data collection tool.**

Closed-ended data: Interviewer frames question and answer options. Closed-ended data is typically collected to confirm (or discard) hypotheses. Closed-ended data is usually generated through such tools as surveys, check lists, structured interviews, and safety audits.
Open-ended data: Interviewer frames topics/aspects of topics or key questions, but leaves open exact wording and answer options. Open-ended data is typically collected to explore/generate new hypotheses. Open-ended data tends to be generated in conversations, focus group discussions, other group discussions, open interviews, biographic interviews.

For tools not listed above, check the evaluation report to assess whether they yield preponderantly closed-ended or open-ended data. If no information = unclear.

- If data collection tool produced predominantly closed-ended data, write "1" in second column.
- If data collection tool produced predominantly open-ended data, write "0" in second column.
- If unclear, write "9".

Save each completed WORD document under its original name.

4. Use **EXCEL sheet C –** [coder-specific]**, in particular the columns** "open-ended" and "closed-ended".
5. Study the **conclusions in the executive summary of the actual evaluation reports**. Take into account the importance each type of data collection tool has had for the conclusions, and whether the tools have produced preponderantly open-ended or closed-ended data. How many conclusions are *based on open-ended data*, *based on closed-ended* data?

Write the results the respective columns, coding them as follows:

- Almost all conclusions= 4
- Most = 3
- Some = 2
- Almost none = 1

Save the EXCEL file under its original name.

Coding TOR/ evaluation reports: beneficiaries, objectives, theory of change

1. Use **EXCEL sheet C –** [coder-specific], in particular the columns "def. of beneficiaries", "objectives", and "TOC".

2. Check the evaluation Terms of Reference (TOR) to answer the question: Is the population that is supposed to benefit from the intervention well defined?

- ***Well defined* (score = 2)** includes information on sex, age group, geographical region and number of beneficiaries.
- ***Adequately defined* (score = 1)** includes information on sex and number of beneficiaries.
- ***Poorly/ not defined* (score = 0)** includes information on number or sex only, or no information at all.

Write the respective scores in the column "def. of beneficiaries".

3. Check the TOR to answer the question: Are the objectives or expected outcomes of the intervention stated in clear terms?

The objectives or expected outcomes are:

***Well defined* (score = 2)** if they fully meet all of the following characteristics: specific, measurable, and attainable.

- **Specific**: They state in reasonably precise terms what change the intervention is supposed to accomplish or contribute to *(for instance, 'improved access to legal aid' – not 'a life free of violence')*
- **Measurable**: They come with indicators for qualitative or quantitative measurement of progress
- **Attainable**: they can realistically be achieved with the resources and time-frame of the project *(not 'a life free of violence')*

***Vaguely defined* (score = 1)** if they have 2 of the characteristics listed above.

***Poorly defined* (score = 0)** if they have up to 1 of the characteristics.

Write the respective numbers in the column "objectives".

4. Check the TOR to answer the question: Is the way in which the intended objectives or outcomes will be achieved (theory of change, TOC) described in a coherent manner?

- ***Coherent*** if cause-to-effect links are clearly described and ***factors external to the intervention*** (context) are taken into account (score = 2).
- ***Deficient*** if cause-to-effect links are described ***without taking into account context factors*** (score = 1).
- ***Absent*** if there is ***no explanation*** of the links between activities and intended outcomes/ objectives (score = 0).

Write the respective scores in the column "TOC" and save the EXCEL file under its original name.

C. Google search

This task consists in searching all 39 evaluation reports on Google and determining how often the respective report appears.

1. Search terms for each evaluation: (a) name of the funding organisation, (b) title of the evaluation as stated on the evaluation report and (c) type of evaluation as stated in the report, such as "mid-term evaluation", "external evaluation", "review", if the title does not include any such term.

2. Use google search. Enter every single word of the search terms in quotation marks and add AND between the words. (The quotation marks stop google from displaying "synonyms" and the AND makes sure it will only display results that contain all search terms.)

3. Note down the resulting number in **EXCEL sheet A –** [coder-specific] in the column "citations".

Save the EXCEL file under its original name.

D. Recoding 2nd round results on bias

This task can only be performed when all coders assessing bias will have completed that task. These assessments will be provided in a dedicated sub-folder of your dropbox folder.

You will work on (1) bias assessments provided by other coders in the 2nd coding round, as well as (2) 1st coding round reporting sheets.

To record your answers, use **EXCEL sheet A –** [coder-specific], which includes the columns "selection bias", "justification for selection bias", "power bias" and "justification for power bias".

1. Read **both bias assessments** for each evaluation report listed on the EXCEL sheet. *(The bias assessments will be made available to you in a dedicated dropbox sub-folder as soon as we receive them from the other coders.)*

   - The stakeholder bias assessments list the different data collection methods and the groups of informants that correspond with each data collection method. They determine what types of stakeholders (active intervention stakeholders, passive intervention stakeholders or non-stakeholders) these groups of informants represent, and assess the potential stakeholder bias – defined in the text box below.
   - The power bias assessments determine power bias – defined below – in a similar way.

2. Check **#36, 37 and 38 of the respective 1st coding round reporting sheet.** These questions assess the importance of a given type of stakeholders for the conclusions that the respective evaluation report has drawn from the evidence.

KEY DEFINITIONS

Selection bias is defined as follows: We define **selection bias** as problems related to the objectivity and fairness of data collection which are caused by the way in which people who provide information (informants) are chosen.

Selection bias is introduced if (1a) only active intervention stakeholders are informants, and/or (1b) informants are selected exclusively by active intervention stakeholders. Such bias can be prevented if substantial information is gathered from informants who are (2a) randomly selected or (2b) selected by the evaluation team according to a scheme which factors in their different relationships to the evaluated intervention.

Power bias is defined as follows: We define **power bias** as problems related to the objectivity and fairness of data collection which are caused by placing people who provide information (informants) into situations where unequal power relationships between those present make it difficult for some or all participants to express themselves freely.

For instance, in societies where power dynamics expect women to remain silent or to avoid publicly contradicting men, focus groups bringing together women and men would induce power bias. Similar risks exist when beneficiaries are interviewed by staff of the implementing organisation, or in the presence of such staff. Power dynamics bias is prevented if (4a) respondents may provide information anonymously.

3. Step 1 (previous page) has provided selection bias assessments for each group of informants, and determined which types of stakeholders these groups of informants represent. Thus you know how "selection biased" information from a particular type of stakeholders is.

   Step 2, based on tasks #36, 37 and 38 of the reporting sheets, is about the respective importance of different types of stakeholders for the conclusions in the evaluation report.

   On the basis of the information gathered in those two steps, judge how much selection bias is potentially present in the information base for the overall conclusions of the evaluation report.

   Is selection bias:

   - Absent/very weak (0)
   - Weak/rather weak (1)
   - Rather strong/strong (2) or
   - Very strong (3)?

   Write the respective number in the column "selection bias" of excel file A.

   In the next column ("justification for selection bias"), **justify your assessment in up to 100 words.**

4. **Repeat the exercise for power bias**. Is **power bias**: it absent/very weak (0), weak/rather weak (1), rather strong/strong (2) or very strong (3)?

   Write the respective number in the column "power bias" in EXCEL sheet A - [coder-specific]. **Justify your assessment in up to 100 words** in the column "justification for power bias".

Work package 3:

Overview

Three types of tasks have been prepared for you:

7. Re-coding specific aspects for [coder-specific] reports you have worked on in the 1st coding round
8. Entering data from the 1st coding round reporting sheets
9. Coding aspects of other coders' 2nd round reports. (That task will have to come after all other coders will have completed their coding work.)

We have foreseen a total of [coder-specific] full working hours for the completion of these tasks.

You have been invited to a dropbox folder which includes the following items:

- The 39 evaluation reports in our QCA set
- The 1st coding round reporting sheets for those 39 reports
- A separate folder with TOR or similar documents for evaluation reports that do not include the TOR in the actual report
- EXCEL tables that you will need for coding
- A sub-folder where to store your work

Please call Wolf or send us an e-mail message (Wolf cc. Michaela) whenever any question arises. You are invited to **store all work-in-progress for this assignment in our shared dropbox folde**r so that we can monitor it and get in touch with you in case anything needs to be corrected.

A. Recoding 'your' 1st round reports

The following three tasks are to be performed on the following reports only: [coder-specific]. Please record all data on the EXCEL sheets provided in 'your' dropbox folder.

Coding evaluation reports: data used in conclusions

1. For each of the evaluation reports listed above, study the conclusions stated in the executive summary. Please assess: When looking at the conclusions stated in the executive summary, how many are…

- based on ***original data*** (collected as part of the evaluation with data collection tools provided by the evaluator(s)).
- based on ***original analysis*** (data collected by others with data collection tools provided by others, analysed by the evaluator(s)).
- based on ***documents*** (data analysis cited by the evaluator(s)).

2. Code your assessment in the respective columns of EXCEL sheet C – [coder-specific] ("original data", "original analysis" and "documents") as follows:

- Almost all conclusions = 4
- Most conclusions = 3
- Some conclusions = 2
- Almost none = 1

Coding 1st round reporting sheets: types of data

6. Create a **WORD document** for each evaluation report in **"EXCEL sheet C –**[coder-specific]**"**. Name it in the following way: "[Last name of main author] [year of publication] – method – [your first name]". Take the information for last name of main author and year of publication from the EXCEL sheet C. For instance, "Chibuta 2011 – method – Sanja.doc".

7. In each Word document, create a table with two columns. Name the columns “Data collection tool” and “Type of data” respectively.

3. For each of the [coder-specific] evaluations, study **question 25 in the 1st coding round reporting sheets for the respective evaluation**. Please determine, for every data collection tool, whether the tool has provided preponderantly closed-ended data or open-ended data. Create a new line in the document for each data collection tool.

- ***Closed-ended data***: Interviewer frames question and answer options. Closed-ended data is typically collected to confirm (or discard) hypotheses. Closed-ended data is usually generated through such tools as surveys, check lists, structured interviews, and safety audits.
- ***Open-ended data***: Interviewer frames topics/aspects of topics or key questions, but leaves open exact wording and answer options. Open-ended data is typically collected to explore/generate new hypotheses. Open-ended data tends to be generated in conversations, focus group discussions, other group discussions, open interviews, biographic interviews.

For tools not listed above, check the evaluation report to assess whether they yield preponderantly closed-ended or open-ended data. If no information = unclear.

- If data collection tool produced predominantly closed-ended data, write “1” in second column.
- If data collection tool produced predominantly open-ended data, write “0” in second column.
- If unclear, write “9”.

Save each completed WORD document under its original name.

3. Use **EXCEL sheet C –** [coder-specific]**, in particular the columns** “open-ended” and “closed-ended”.

4. Study the **conclusions in the executive summary of the actual evaluation reports**. Take into account the importance each type of data collection tool has had for the conclusions, and whether the tools have produced preponderantly open-ended or closed-ended data. How many conclusions are *based on open-ended data*, *based on closed-ended* data?

Write the results the respective columns, coding them as follows:

- Almost all conclusions= 4
- Most = 3
- Some = 2
- Almost none = 1

Save the EXCEL file under its original name.

Coding TOR/ evaluation reports: beneficiaries, objectives, theory of change

1. Use **EXCEL sheet C –** [coder-specific], in particular the columns “def. of beneficiaries”, “objectives”, and “TOC”.

2. Check the evaluation Terms of Reference (TOR) to answer the question: Is the population that is supposed to benefit from the intervention well defined?

- ***Well defined* (score = 2)** includes information on sex, age group, geographical region and number of beneficiaries.
- ***Adequately defined* (score = 1)** includes information on sex and number of beneficiaries.
- ***Poorly/ not defined* (score = 0)** includes information on number or sex only, or no information at all.

Write the respective scores in the column "def. of beneficiaries".

3. Check the TOR to answer the question: Are the objectives or expected outcomes of the intervention stated in clear terms?

The objectives or expected outcomes are:

***Well defined* (score = 2)** if they fully meet all of the following characteristics: specific, measurable, and attainable.

- **Specific**: They state in reasonably precise terms what change the intervention is supposed to accomplish or contribute to *(for instance, 'improved access to legal aid' – not 'a life free of violence')*
- **Measurable**: They come with indicators for qualitative or quantitative measurement of progress
- **Attainable**: they can realistically be achieved with the resources and time-frame of the project *(not 'a life free of violence')*

***Vaguely defined* (score = 1)** if they have 2 of the characteristics listed above.

***Poorly defined* (score = 0)** if they have up to 1 of the characteristics.

Write the respective numbers in the column "objectives".

4. Check the TOR to answer the question: Is the way in which the intended objectives or outcomes will be achieved (theory of change, TOC) described in a coherent manner?

- ***Coherent*** if cause-to-effect links are clearly described and ***factors external to the intervention*** (context) are taken into account (score = 2).
- ***Deficient*** if cause-to-effect links are described ***without taking into account context factors*** (score = 1).
- ***Absent*** if there is ***no explanation*** of the links between activities and intended outcomes/ objectives (score = 0).

Write the respective scores in the column "TOC" and save the EXCEL file under its original name.

B. Entering data from the 1st coding round

This task needs to be performed with all 39 reporting sheets from the 1st coding round.

1. Open **EXCEL Sheet A -** [coder-specific] (provided to you in the dropbox folder).
2. Enter the data for **reporting sheet task number 2, 3, 4, 5, 6, 33, 35, 36, 37, 38, 43, 44 of all** evaluations in the respective columns. Write down only numbers, as follows:
    - If answer options yes/no: yes = 1, no = 0.
    - If answer options yes/rather yes etc.:

· yes = 4
· rather yes = 3
· rather not = 2
· no = 1
· not sure = 9.

- If answer options almost all/most etc.:
  · almost all = 4
  · most = 3
  · some = 2
  · almost none = 1.
- If no answer provided = 99.

3. **Review task number 12 of the reporting sheets of all** evaluations. If a single option is marked, code the answer in Excel file A – Sanja, column "evaluator independence", as follows:

   - A = 0
   - B or C = 2
   - D = 4.

   If question 12 has no answer or more than one option is marked, check the ToR of the respective evaluation. If ToR is not available, check the report and code your assessment as defined above. If you cannot determine which answer option to choose, code "9".

4. Review **task number 46 of the reporting sheets** of all evaluations. If the reporting sheet does not state any reference to gender issues, enter the code 0 in column "gender issues" of Excel Sheet A - Sanja. For all reporting sheets which state that gender issues are mentioned, enter the code 1.

5. Review **task number 64** of the reporting sheets of all evaluations.

   - If question 64 is answered with "yes", write 1 in column "respectful vocabulary" of Excel Sheet A - Sanja.
   - If question 64 is answered with "no", write 0.
   - If there is no answer, then read the executive summary and the first 10 pages of the report and use your own judgment to assess whether the presentation is respectful and free of gender bias, or not.

6. Save the EXCEL sheet under its original name.

C. Recoding 2nd round results on bias

This task can only be performed when all coders assessing bias will have completed that task. These assessments will be provided in a dedicated sub-folder of our shared dropbox folder.

You will work on (1) bias assessments provided by other coders in the 2nd coding round, as well as (2) 1st coding round reporting sheets.

To record your answers, use **EXCEL sheet A –** [coder-specific], which includes the columns "selection bias", "justification for selection bias", "power bias" and "justification for power bias".

5. Read **both bias assessments** for each evaluation report listed on the EXCEL sheet. *(The bias assessments will be made available to you in a dedicated dropbox sub-folder as soon as we receive them from the other coders.)*

- The stakeholder bias assessments list the different data collection methods and the groups of informants that correspond with each data collection method. They determine what types of stakeholders (active intervention stakeholders, passive intervention stakeholders or non-stakeholders) these groups of informants represent, and assess the potential stakeholder bias – defined in the text box below.
- The power bias assessments determine power bias – defined below – in a similar way.

6. Check **#36, 37 and 38 of the respective 1st coding round reporting sheet.** These questions assess the importance of a given type of stakeholders for the conclusions that the respective evaluation report has drawn from the evidence.

KEY DEFINITIONS

Selection bias is defined as follows: We define **selection bias** as problems related to the objectivity and fairness of data collection which are caused by the way in which people who provide information (informants) are chosen.

Selection bias is introduced if (1a) only active intervention stakeholders are informants, and/or (1b) informants are selected exclusively by active intervention stakeholders. Such bias can be prevented if substantial information is gathered from informants who are (2a) randomly selected or (2b) selected by the evaluation team according to a scheme which factors in their different relationships to the evaluated intervention.

Power bias is defined as follows: We define **power bias** as problems related to the objectivity and fairness of data collection which are caused by placing people who provide information (informants) into situations where unequal power relationships between those present make it difficult for some or all participants to express themselves freely.

For instance, in societies where power dynamics expect women to remain silent or to avoid publicly contradicting men, focus groups bringing together women and men would induce power bias. Similar risks exist when beneficiaries are interviewed by staff of the implementing organisation, or in the presence of such staff. Power dynamics bias is prevented if (4a) respondents may provide information anonymously.

7. Step 1 (previous page) has provided selection bias assessments for each group of informants, and determined which types of stakeholders these groups of informants represent. Thus you know how "selection biased" information from a particular type of stakeholders is.

   Step 2, based on tasks #36, 37 and 38 of the reporting sheets, is about the respective importance of different types of stakeholders for the conclusions in the evaluation report.

On the basis of the information gathered in those two steps, judge how much selection bias is potentially present in the information base for the overall conclusions of the evaluation report.

Is selection bias:

- Absent/very weak (0)
- Weak/rather weak (1)
- Rather strong/strong (2) or
- Very strong (3)?

Write the respective number in the column "selection bias" of excel file A.

In the next column ("justification for selection bias"), **justify your assessment in up to 100 words.**

8. **Repeat the exercise for power bias**. Is **power bias**: it absent/very weak (0), weak/rather weak (1), rather strong/strong (2) or very strong (3)?

Write the respective number in the column "power bias" in EXCEL sheet A - [coder-specific].
Justify your assessment in up to 100 words in the column "justification for power bias"

Annex IX: Guidelines for interviews on evaluation effects

Preliminaries

Thank you for making time for this interview. I am keen on talking to you because I would like to find out about the effects the evaluation of XYZ in XXXX has produced. **We are interested in the use and effects of evaluations, to find out what approaches and methods work under what kinds of circumstances. Our purpose is <u>not</u> to judge** who has made the best evaluation, but really to learn from a wide range of approaches and situations.

Can I record the interview? I will not share the recording with anyone. If you say anything that we could use as a quote for our report and it is clear the quote can only come from you, then we will ask for your permission first, to make sure the quote is exact.

Do you have any questions before I start asking questions?

According to our information,

[**Evaluators**:] you have evaluated the project/ programme/ organization/ initiative.

[Implementers:] you have been involved in the implementation of the evaluated project/ programme/ organization/ initiative.

[Funders:] your organization has financed the project/ programme/ organization/ initiative that was evaluated.

Is that correct? *[No need to ask that if it is totally clear, e.g. where the evaluator's name is on the report as the author's.]*

We have planned up to 45 minutes for this interview. Is that OK for you?

Questions on evaluation effects.

We believe that any evaluation produces effects. Both positive and negative effects are interesting.

Usually, an evaluation starts with a preparatory phase, then the evaluation is carried out, and finally the report is shared or published. Different people are involved in the evaluation: those who have implemented the project or other intervention, those who have funded it, the wider development community and the ultimate beneficiaries of the project, programme or intervention.

[Evaluators:] As the evaluator/one of those who have evaluated the project/ programme/ organization/ initiative,

[Implementers:] As someone who has been involved in the implementation of the evaluated project/ programme/ organization/ initiative,

[Funders:] As the representative of a donor to the project/ programme/ organization/ initiative,

Could you describe to me how these different groups of people (project implementers, donors, beneficiaries) were involved in the different phases of the evaluation?

1. Let's start with the **preparation of the evaluation. Who was involved, and how?**
2. How about the **implementation of the evaluation. Who was involved, and how?**

3. When you think of the preparation and the implementation of the evaluation, did you notice any **effects on the different groups of people involved? Both positive and negative** effects are interesting to us.
4. Finally, **who read or used the evaluation report?**
5. **Have there been any effects – positive or negative - on one or several of the four groups after the evaluation report was published?** Again, the four groups are: (1) those who have implemented the project/intervention, (2) those who have funded it, (3) the wider development community and (4) the ultimate beneficiaries of the project.
6. **Are you aware of any people who are not directly involved in the project/ programme** who have used the evaluation report, or who might have experienced some – **positive or negative – changes because of the report?**

[If interviewee asks whether a particular effect would also count, always say yes. In general: Always ask for a concrete example; a link to observable behavior (activities, decisions).]

If respondents asks: Additional explanations for different effect types. Effects on…

those who have implemented the project, programme or other activity that was evaluated: *For example, have the evaluation process and the report helped them improve their work or any related activities?*

those who have financed the project, programme or other activity that was evaluated: *For example, how has the evaluation contributed to their decision making, to their advocacy work, or to their accountability towards others.*

the wider development community*: How has the evaluation contributed to learning beyond the project/ programme/ other activity that was evaluated?*

the intended beneficiaries or "ultimate target groups" of the project/ programme or other activity that was evaluated: *These are the women and men, girls and boys whose life is supposed to be improved by the project or programme. Has the evaluation had any good or bad effects on them, directly or indirectly?*

7. **Are there any other effects that come to your mind?**
8. **In your opinion, what factors have made it easier** – or harder! – for the evaluation to produce the effects that you have described?

Questions on information gaps.

See "summaries.xlsx", items 12 and 13.

[If evaluation not found on the web.] Has the evaluation been published? **If not, do you know why?**

Questions on contact information of others.

We plan to run an on-line survey with a member of the evaluation team/ the donor/ someone from the organisation(s) that implemented the project/ programme to find out about their perceptions on the effect. It has been difficult to obtain contact details. **Would you have the e-mail addresses?**

Announcement of survey.

In the coming weeks, we will distribute the survey. We will use the information you have provided us with to develop the questionnaire. But it would be great if you could participate in the survey as well, as the questions and answers will be more standardised.

If you cannot, would you have a colleague in your evaluation team / organisation who could take the survey?

This has been a very useful interview. Many thanks for your kind support!

Annex X: Guidelines for process tracing interviews

[Two different guidelines – for implementers and funders of the evaluated intervention and for evaluators – content very similar though. Here only reproduced guidelines for interviews with implementers and funders of the intervention evaluauted…]

Preliminaries (5 minutes)

Thank you for making time for this interview. As announced in my earlier e-mail message, we have selected the **evaluation *[name, country, author, year]*** for process training.

We have planned interviews with three persons who are knowledgeable about the evaluation **to find out about the effects the evaluation has produced, and how – in the interviewee's opinions – these effects have come about**.

The ultimate **purpose** is to understand what can be done in evaluations to make them most useful and effective. So, this interview is really about the evaluation – not about the intervention that has been evaluated.

Can I record the interview? I will not share the recording with anyone; this interview is **confidential**. If you say anything that we could use as a quote for our report and it is clear the quote can only come from you, then we will ask for your permission first, to make sure the quote is exact.

We have planned **45-60 minutes** for this interview. Is that feasible for you?

In case the **phone line breaks down, please wait for me to call you back**.

Do you have any questions before I start asking questions?

According to our information,

[Implementers:] you were involved in the implementation of the evaluated project/ programme/ organization/ initiative.

[Funders:] your organization funded the project/ programme/ organization/ initiative that was evaluated.

Is that correct? *[No need to ask that if it is totally clear, e.g. where the evaluator's name is on the report as the author's.]*

Question on evaluation effects (5-10 minutes)

1. According to the survey responses we have received, **this evaluation has produced the following effects:** *[Read out effects as per QCA steps, use definitions below]*

Definitions for effects

Action effects . The evaluation has helped to **change, or to reinforce, the way** an intervention has been implemented. Such effects can occur at the **level of the evaluated intervention** (such as in a mid-term review), or in a follow-up intervention, and at the level of **wider development practice by those who have implemented the intervention and their funders**.

Persuasion effects. The evaluation has **convinced** others to **support** the intervention that was evaluated (for example, donors maintain or increase their funding) or the policies it advocates for.

Learning effects = Insights and influence that affect the wider development, women's rights and evaluation communities, beyond and independently from the evaluated intervention.

Based on feed-back you may have received, **do you agree with these findings?**

[If interlocutor disagrees] I see. From your perspective, **what changes has the evaluation contributed to –**

- in terms of **supporting decisions** on the future course of the intervention or similar interventions by organisations involved in the intervention ("internal learning")
- in terms of **persuading** others – for example, accountability to donors or advocacy with political decision-makers
- in terms of **wider learning** for organisations and individuals who have not been involved in the intervention

[Probe** to obtain **concrete descriptions of the outcomes]

So, these are the effects that you know about. Now we would like to find out how the effects have come about.

Questions on the causes of the evaluation effects (30-45 minutes)

2. In your opinion, **who did what, or who made what decision**, to bring about each of the effects we just discussed?

3. What do you know as to how **that action/ decision you have described has come about?**

[**Continue **probing** for specific actors and actions - **until interlocutor mentions elements listed as conditions in our QCA.]

***[**Probe to make sure interlocutor describes t**he causal pathway for each effect** mentioned.]*

4. You have mentioned ***[**reiterate the **conditions that relate closely to what the interlocutor has described].*** There are some other conditions as well that we believe have played an important role in producing effects in the case of your evaluation *[name **conditions as described in QCA "pathway"** for the evaluation, for instance, "according to our measurement your evaluation was very good on communication"].* What do you think about this? ***[If extra interview time available]***

5. **What decisions or actions, by whom, did each of these conditions bring about?** We are interested in all phases of the evaluation, including also preparation and follow-up.

[Probe for each key condition the interlocutor has mentioned.]

*[If the interlocutor strays from the conditions listed in the model, dwelling on potential conditions we have found to be of very limited relevance, do not explore further – **UNLESS she/ he mentions factors that we have not yet taken into account.]***

6. Is there **anything that you would like to add?**

Many thanks for sharing these details. This helps enormously in understanding whether and how evaluations in this field can cause effects.

Practical issues

[Only for evaluation reports that have not been published.]

I have a small practical question. We plan to make short, one-page **descriptions** of some exemplary evaluations, including this one. The descriptions are about evaluations that we see as helpful examples that others can used for guidance when designing evaluations in the field of violence against women and girls.

Is it correct that the evaluation has not been published?

[YES →] OK. Then we would suggest we draft the short description and send it for clearance to the organisation that has commissioned the evaluation. Would you that be you, or should we send it to anybody else?

[NO →] OK. Then we erroneously listed the report as unpublished, or maybe the version we have is not the published one. Then I suppose it is OK we go ahead and write these short descriptions.

This has been a very useful interview. Many thanks for your kind support!

Annex XI: Review Terms of Reference

Note: Slight changes to the format have been introduced for visual coherence with the overall report.

Review of evaluation approaches and methods for violence against women and girls interventions

Purpose and objectives

1. The purpose of this review is to improve the international community's knowledge and understanding of the approaches and methods used to evaluate interventions addressing violence against women and girls (VAWG). The review will assess the strengths, weaknesses and appropriateness of these approaches and methods, considering the attributes of the interventions, the contexts in which they take place, and the evaluation questions asked, and identify lessons learned for the improvement of future evaluations on this issue. The review will contribute to global attempts to tackle violence against women and girls by equipping practitioners, including women's rights organisations, with the knowledge they need to assess what works and what doesn't work in VAWG programming, and increasing the number of high quality evaluations of VAWG programmes. This will provide a more solid evidence base for more effective programme design and implementation in the future, contributing to better access to and quality of services for women and girls affected by violence, the reduction and ultimately the elimination of violence against women and girls.

2. Working to eliminate violence against women and girls is a strategic priority for DFID.[17] It was also the focus of the 2013 United Nations Commission on the Status of Women.

3. 2013 is the year when the UK takes the Presidency of the G8. The Foreign Secretary's initiative on preventing sexual violence in conflict is a concrete example of the commitment the UK is showing in this area. The UK will use its Presidency of the G8 to ask some of the world's most powerful nations to make new commitments to help shatter the culture of impunity for those who rape in warzones, to increase the number of successful prosecutions and to help other nations build stronger national capabilities to end the suffering caused by this violence.

4. The review will therefore be a timely addition to research on violence against women and girls and has the potential to improve programming by donor agencies and other relevant actors on this issue.

5. The review will be distinct from a systematic review[18] in that it will focus primarily on evaluations, it is unlikely to have such strict inclusion criteria related to quality and it will be principally interested in the evaluation questions, designs and methods used, in addition to evaluation findings. It will aim to complement and build on existing literature around evaluations of violence against women and girls interventions.[19]

6. The review will be a research product for policy makers, programme staff, evaluators and evaluation commissioners in the international development community to improve evaluations of VAWG programming globally. The review will seek to engage with key players in the sector to encourage maximum communication and uptake of the review's findings, including UN agencies, bilateral agencies,

[17] DFID's Strategic Vision for Girls and Women includes four pillars: delay first pregnancy and support safe childbirth, economic assets direct to women and girls, get girls through secondary school, prevent violence against girls and women. See Annex 1: DFID Strategic Vision for Girls and Women and Annex 2: DFID pillar on Violence Against Women and Girls for more information

[18] DFID How To Note on Assessing the Strength of Evidence, p.8: 'Systematic Review designs adopt systematic methods to searching for literature on a given topic. They interrogate multiple databases and search bibliographies for references. They screen the studies identified for relevance, appraise for quality (on the basis of the research design and methods they employ), and synthesise the findings using formal quantitative or qualitative methods. Systematic Reviews are always clearly labelled as such.18 They represent a robust, high quality technique for evidence synthesis.' https://www.gov.uk/government/publications/how-to-note-assessing-the-strength-of-evidence

[19] We do not expect the review's budget to exceed that of a DFID systematic review.

civil society organisations, women's rights organisations and private sector foundations.[20] To this end, the contractors will be required to detail an innovative communications strategy for the review, which can include, for example, the presentation of its findings at international evaluation and/or gender-related events (findings from Stage 1 of the review could be presented at the Commission on the Status of Women 2014), workshops to reach a range or audiences including women's rights organisations, publication in a relevant journal and the use of digital media.

7. The DFID technical lead for this piece of work will be Evaluation Department Evaluation Officer/Evaluation Department DESA Social Development Advisor, supported by a Social Development Evaluation Advisor (Evaluation Department). The review will receive further support from DFID's conflict and humanitarian department (CHASE), including the DFID policy lead on violence against women and girls and a conflict and humanitarian evaluation adviser, who will also form part of the team selecting the preferred bidder and participate in the reference group.

Background

8. The international development community is working to tackle violence against women and girls by empowering women and girls, addressing social norms that lead to violence, including working with men and boys, building political will and legal and institutional capacity, and providing comprehensive services to women and girls affected by violence.[21]

9. To date, there has not been a thorough review of evaluation literature in this area. A short Governance and Social Development Research Council (GSDRC) Helpdesk report on evaluation of VAWG programmes found that despite the growing number of interventions the quality of existing evaluations is variable. These evaluations also often assess specific projects rather than wider programmes.[22] The DFID How to Note on Monitoring and Evaluation of VAWG programmes found that, in addition, most evaluations are process rather than impact evaluations. 'This is due to many factors such as the difficulty of obtaining reliable data, the complexity and context-specificity of Violence against Women and Girls interventions, and the political and social dynamics surrounding these issues.'[23]

Scope

10. Stage 1: Scoping of the evaluation landscape (Output 1- scoping report):
    A comprehensive scoping of the landscape of evaluation practice across interventions to tackle violence against women and girls interventions – eg the number, quality, types and range of evaluations, and thematic and geographic coverage, ie. what kind of programmes they cover and where they are based, and whether certain types of programmes have been evaluated more or less.

    The scoping report should also consider unpublished evaluations,[24] if these can be accessed, and if feasible make an assessment of what proportion of evaluations go unpublished and if this has any link with whether findings are negative or positive.

    This stage will be characterised by a systematic search across English-language evaluations and related documentation on VAWG programmes (both programmes where addressing VAWG is the primary focus and other programmes where a VAWG element is integrated, where an intended outcome is to reduce

[20] See Annex 3 for list of key organisations. This is indicative and not a definitive list.
[21] See Annex 4 for DFID's Theory of Change for Violence Against Women and Girls interventions, which details which kind of interventions may come under these four areas.
[22] Governance and Social Development Research Council (GSDRC) Helpdesk Query: Evaluation of programmes relating to violence against women and girls, http://www.gsdrc.org/go/display&type=Helpdesk&id=853
[23] How To Note: Guidance on Monitoring and Evaluation of Programming on Violence Against Women and Girls
[24] DFID's definition of an evaluation includes the need for transparency; therefore they need to be published. However, for the purpose of this review, it will be useful to consider how many unpublished "evaluations" the team are aware of.

VAWG or where the evaluation measures the programme's impact on VAWG, whether intended or unintended) by public, private and not-for-profit actors (including women's rights organisations) in developing countries, including humanitarian contexts/peace and security interventions, interventions in conflict and post-conflict settings, peacekeeping interventions, HIV/AIDS interventions, work on masculinities and social norms, although interventions in developed countries will also be considered. **A proposed methodology for the search, including search terms and a clear typology of the types of interventions of interest (eg. humanitarian interventions, peacekeeping, interventions in conflict and post-conflict settings, HIV/AIDS programmes, work on masculinities and social norms etc), will be developed by potential contractors in the bid for the review.** (Output 1 –scoping report)

11. Stage 2: Narrowing the field of enquiry:
    Producing an inception report (Output 2) indicating how stage 3 of the review will be conducted and making proposals regarding inclusion/exclusion criteria for the review, ***for example*** on the basis of:

    - what counts as an evaluation[25] (DFID is not wedded to any particular evaluation approach, design or method; a broad set of evaluations should be considered for this review, eg. experimental, theory-based, case study based, participatory etc., and should include process and impact evaluations)
    - the quality of the evaluation; eg. what counts as high quality; whether to apply exclusion criteria based on quality when we are principally interested in approaches and methods; if yes: what is good enough to include; whether to include lower quality evaluations in stage 3 of the review in order to learn from mistakes and/or if they ask interesting questions or evaluate key areas;
    - considering the findings from Stage 1, what interventions to include (eg. both interventions where violence against women and girls is the main focus or mainstream interventions that may tackle violence indirectly?; interventions where a reduction in violence against women and girls is a secondary, rather than primary, objective?; which thematic areas to include, given that interventions to tackle violence against women and girls cover several areas;
    - Relevance to DFID's portfolio in this area[26]
    - Attention to the wider literature and evidence base, ie. beyond evaluative evidence
    - Relevance to target audience ie. evaluation specialists, evaluation commissioners, and individuals designing interventions with the potential to address violence against women and girls – both in DFID and the wider international development community.
    - Criteria to be used to assess the rigour and appropriateness, strengths and weaknesses of the designs, approaches and methods used to address the evaluation questions.

    The inception report should also include:

    - Proposed format for the review;
    - Initial thoughts on a communication strategy for the final review.

    Once the inception report and inclusion/exclusion criteria have been agreed with DFID the contractors will identify a sub set of evaluations to review in more detail (Output 3).

12. Stage 3: Reviewing evaluation approaches and methods:
    A review of the evaluation approaches and methods used to evaluate interventions to tackle violence against women and girls (Output 4, 5) – including:
    - *The evaluation purpose and objectives:*

[25] DFID uses the OECD Development Assistance Committee definition of evaluation:
"The systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation, and results in relation to specified evaluation criteria."
[26] See Annex 2 for a summary of DFID's programming under the violence against women and girls pillar of DFID's Strategic Vision for Girls and Women.

- The evaluation purposes and questions posed (perhaps clustered by DAC criteria);
- *The design, approach, methods and tools used:*
  - An assessment of the rigour and appropriateness, strengths and weaknesses of the designs, approaches and methods used to address those evaluation questions, including data collection and sampling methods, considering programme attributes and the context in which it takes place
  - The use of theory of change or where there is no theory of change, how assumptions are dealt with in the evaluation
  - How and when evaluations have been able to assess impact (including what proportion of impact evaluations claim attribution and how many contribution), particularly whether impact evaluations have been able to distinguish between theory failure and implementation failure
  - How evaluations have assessed different types of intervention, eg. how changing individual attitudes and social norms around gender and violence against women and girls have been measured
- *The evaluation process:*
  - Constraints and challenges faced by evaluators (eg. identification of outcome indicators, level of data collection, use of control groups, under-reporting of VAWG, poor access to target group, short time frames for programming, difficulty of systematising qualitative data, novelty of/unfamiliarity with social norm change evaluation tools, confusion about divide/links between prevention and response, lack of capacity/experience of evaluators on VAWG, etc), including how risks were mitigated and challenges overcome; challenges particular to conducting evaluations of VAWG programmes in humanitarian settings, eg. availability of analytical and baseline data, constraints due to the insecurity of the environment
  - Attention to the process of the evaluation, including. ethical and safety considerations, including how evaluators dealt with collecting data from children and vulnerable people on this sensitive subject, how evaluators addressed bias, if evaluations have been participatory and empowering, integrated human rights and included sound gender and power analysis in their approaches and methods
- *Trends and gaps:*
  - Trends – eg. in relation to evaluation questions, designs, approaches and methods (including any emerging methodological preferences, and how these vary by donor)
  - Any gaps, ether in terms of under-evaluated types of programmes, evaluation coverage (including both thematic and geographic coverage – what types of interventions have been evaluated and where these interventions have taken place*),* neglected evaluation questions and under-used approaches and methods
- *Use and usability:*
  - Attention to the use of evaluations, eg. whether there is evidence that evaluation findings are being used for programming and practice, policy change and advocacy
- *Recommendations* to strengthen the quality, value and use of future evaluations of VAWG programming, eg in relation to important evaluation questions, designs, approaches and methods
- A short summary of the key evaluation findings, perhaps shown in a table together with the methods used, should also be included in the review.

In addition to this review of the evaluations themselves the report will also need to include sections on:

- Communication strategy for the review, considering a range of key stakeholders in the international development community, including women's rights organisations, and identifying options for publication and presentation at relevant evaluation and thematic events, as well as additional specialist papers for submission to relevant journals.

13. Stage 4: Communicating the findings:
    - Presentation of the findings to DFID staff
    - Presentation of the findings at international evaluation and/or gender-related events
    - Workshop with women's rights organisations

- In a relevant journal
- Digital communications
- The team may be required to produce additional specialist papers for submission to journals and presentation at conferences. This will be agreed once the team's communication strategy is approved.

14. Users of the review

Organisations and agencies working on VAWG including women's rights groups, and with a particular focus on the following:

- Evaluation commissioners
- Evaluators
- Programme staff and policymakers working on Violence Against Women and Girls

15. Outputs

1. **Output 1: Scoping report** (with possibility to present to DFID staff) detailing results of evaluation scoping exercise.

   Final report to be submitted ten weeks after the start of the contract

2. **Output 2:** An **inception report** (draft and final) for the review, which will include the inclusion criteria for the review, the proposed format and initial thoughts on a communication strategy for the final review.

   Final inception report to be submitted sixteen weeks after the start of the contract

3. **Output 3: List of the proposed evaluations to be included** in the detailed review, using the agreed above inclusion/exclusion criteria.

4. **Output 4:** A **draft of the review** addressing all elements outlined in the scope above. The review should use tables, graphs and other visuals where appropriate to present information in an accessible way.

5. **Output 5:** The **final review** of no more than 30 pages (excluding annexes), following suggestions and revisions to the draft report.

   To be completed nine months after the start of the contract.

6. **Output 6:** Two seminars (Output 6) including PowerPoint presentation on principal and interesting findings
   i. For key stakeholders in the international development community, including women's rights organisations and the VAWG Research and Innovation Fund consortia and VAWG Helpdesk organisations. This will be hosted either by DFID or by another member of the Reference Group
   ii. For DFID staff

   To be completed by ten months after the start of the contract.

7. **Output 7**: The team may be required to produce additional specialist papers for submission to journals and presentation at conferences. This will be agreed once the team's communication strategy is approved.

16. Methods

The contractors may want to employ any or all of the following methods for the review. Any sampling method used to select evaluations for inclusion in the stage 3 review will need to be developed by the contractors and agreed by DFID:

- Desk-based research

- Grey literature review
- Evidence mapping
- Quantitative and qualitative analysis of the evaluation material
- Qualitative interviews with relevant stakeholders, including, but not limited to, bilateral and multilateral donors, CSOs, including women's rights organisations, and private sector foundations.
- Stakeholder mapping
- Stakeholder survey

17. Requisite skills and knowledge

The team will need to demonstrate the following:

- Knowledge of and experience of using a wide range of evaluation approaches and methods, both qualitative and quantitative;
- Methodological openness;
- Knowledge of gender issues related to development and extensive experience in gender analysis
- Knowledge of violence against women and girls, associated interventions, and research and evaluation literature
- Extensive experience of evaluation, both quantitative and qualitative
- Excellent writing and communication skills, including ability to present information in a range of visual ways to increase its impact

18. External reference group

DFID will set up an external reference group of key stakeholders which will bring international expertise on a diverse range of evaluation approaches and methods, experience in gender-related evaluations, knowledge of violence against women and girls, and particularly of evaluations of VAWG. The reference group is to include representatives from bilateral and multilateral organisations, including UN Women, civil society organisations, including women's rights organisations, and private sector foundations. The group will provide comments on draft outputs and facilitate access to evaluation documentation through their contacts. In order to increase access to and uptake of the review's findings, these will also be communicated to the reference group members' own networks.

19. Timing of the review

We anticipate work to start on the review in August/early September 2013 and finish in May 2014.

20. Indicative Structure of the Review

- Executive summary;
- Methods used in the review; inclusion/exclusion criteria; any limitations of the review;
- Summary of findings from the scoping report
- Analysis of the approaches and methods used to address different evaluation questions (their rigour and appropriateness, strengths and weaknesses); use of theory of change; constraints faced by evaluators;
- Analysis of evaluation process, including ethics and safety considerations
- Identification of trends and any gaps ether in terms of evaluation coverage (geographic or thematic), neglected evaluation questions and under-used evaluation approaches and methods;
- A brief summary of the evaluation findings and lesson learned*;*
- Recommendations to strengthen the quality, value and use of future evaluations, eg. in relation to important evaluation questions, approaches and methods.

21. Guidance on definitions

Agreed Conclusions from the 57th Session of the UN Commission on the Status of women[27]:

[27] CSW, (2013) Agreed Conclusions, United Nations Commission on the Status of Women 2013 http://www.unwomen.org/how-we-work/csw/

Para 10: "The Commission affirms that violence against women and girls is rooted in historical and structural inequality in power relations between women and men, and persists in every country in the world as a pervasive violation of the enjoyment of human rights. Gender-based violence is a form of discrimination that seriously violates and impairs or nullifies the enjoyment by women and girls of all human rights and fundamental freedoms. Violence against women and girls is characterized by the use and abuse of power and control in public and private spheres, and is intrinsically linked with gender stereotypes that underlie and perpetuate such violence, as well as other factors that can increase women's and girls' vulnerability to such violence."

Para 11: "The Commission stresses that "violence against women" means any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women and girls, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life. The Commission also notes the economic and social harm caused by such violence."

DFID uses the OECD Development Assistance Committee **definition of evaluation:**

"The systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation, and results in relation to specified evaluation criteria."

DFID also requires its evaluations to:

- Be transparent, ie. published
- Be independent
- Use a systematic and robust methodology

Quality in evaluation: The study team will be required to develop criteria for assessing the quality of evaluations during the inception phase. The following references may be useful:

Shaxson, Louise, (2005) "Is your evidence robust enough? Questions for policymakers and practitioners," *Evidence and Policy: A Journal of Research, Debate and Practice*, v. 1, no. 1.

http://policyimpacttoolkit.squarespace.com/library/is-your-evidence-robust-enough-questions-for-policy-makers-a.html

Chapter 6: Stern et al, (2012) *Broadening the range of designs and methods for impact evaluations* (London: DFID).

https://www.gov.uk/government/.../design-method-impact-eval.pdf

Spencer et al, (2003) *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence* (London: Cabinet Office).

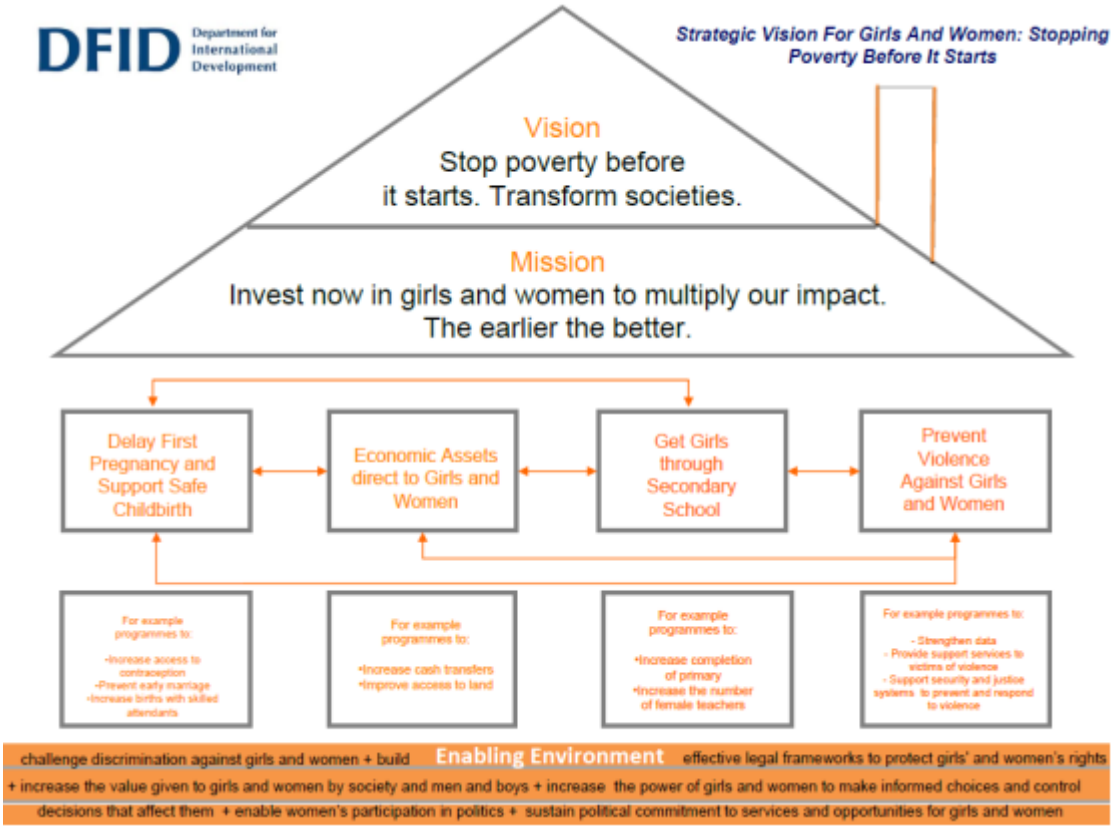
www.civilservice.gov.uk/wp.../a **quality framework** tcm6-7314.pdf

22. Contractual Issues

The project contractor is contracted by DFID and is accountable to DFID. The review will be the intellectual property of DFID. The contractor will report to Clare McCrum (C-McCrum@dfid.gov.uk) on the overall task and to John Murray (J-Murray@dfid.gov.uk) on all contractual matters. The contract will be output-based and payment will be made once outputs are signed off by DFID. DFID will set up and manage a small external reference group to comment on draft reports and provide guidance to support this work. The contractors are obliged to wait until approval is granted from DFID (this will at times include a 2 week period for reference

group comments) before proceeding with the next stages of the review. The report should credit DFID for its contribution to the project. DFID will provide a logo for use in the report.

DFID Strategic Vision for Girls and Women



DFID pillar on Violence Against Women and Girls

Violence against women and girls (VAWG) is the most widespread form of abuse worldwide, affecting one third of all women in their lifetime. Addressing violence against women and girls is a central development goal in its own right, and key to achieving other development outcomes for individual women, their families, communities and nations. DFID's Business Plan (2011-2015) identifies tackling violence against women and girls as a priority and commits DFID to pilot new and innovative approaches to prevent it.[28]

Progress made: The Strategic Vision One Year On[29]:

DFID is scaling up its response on violence against women and girls and there are 20 country offices currently working on violence against women and girls programmes. The majority of these are in fragile and conflict-affected states. There are also another 9 regional or global programmes, for example Programme Partnership Arrangement programmes, Global Girls Research Initiative and the Asia Regional Trafficking Programme.

Recognising that there are still major gaps in the evidence about violence, in November 2012 Secretary of State Justine Greening launched the Violence Against Women and Girls Research and Innovation Fund, which will

[28] How To Note: Guidance on Monitoring and Evaluation for Programming on Violence against Women and Girls

[29] DFID Strategic Vision for Girls and Women: One Year On, 2012 www.dfid.gov.uk/Documents/publications1/StrategicVision-OneYearOn.pdf

invest £25 million over five years 'to drive innovation, generate ground-breaking new evidence and support new prevention programmes. By testing out new approaches and the rigorous evaluation of existing programmes, we can better understand what works in tackling the root causes of violence against women and girls in some of the poorest countries of the world.'[30]

DFID's Violence Against Women and Girls portfolio:

The work of DFID's country programmes cuts across the four themes set out in the VAWG Theory of Change. The themes are:

- Empower Women and Girls
- Change Social Norms
- Build Political Will and Institutional Capacity
- Provide Comprehensive Services

While there are examples of single sector approaches, most programmes work across more than one theme, and can include a range of activities. This work can include initiatives at the policy, service delivery and community levels, and various programmes incorporate an integrated approach which includes working at each of these levels (e.g. the Rights and Governance Challenge Fund in Bangladesh).

In terms of areas of focus, the large majority of programmes focus on institutional strengthening, for example of the security and justice sectors. This includes working with relevant government ministries, the police, army, and informal justice mechanisms. These are followed by programmes aiming to empower women and girls; deliver services; and address social norms. It is important to note that many programmes aim to build institutional capacity to provide services. For example, various programmes reflect increasing interest in strengthening police capacity by establishing support units for victims and survivors of VAWG, and supporting communities to establish safe spaces for women and girls.

The four themes further encompass a range of interventions – such as providing education and skills training to women and girls, undertaking media campaigns, supporting the capacity of government ministries, and providing legal, psychosocial and medical services through safe spaces and protection centres. DFID is scaling up its work to combat VAWG in humanitarian settings and is implementing the UK National Action Plan on UN Security Council Resolution 1325.

List of key organisations

- UN Women
- UN Trust Fund on Ending Violence Against Women and Girls
- UNICEF
- UNFPA
- CIDA
- Danida
- AusAid
- USAID
- Oxfam
- ActionAid
- Womankind Worldwide
- Plan International
- IPPF
- Comic Relief
- Sigrid Rausing Trust
- Oak Foundation
- International Rescue Committee

[30] Justine Greening: Eliminating Violence against Women and Girls http://www.dfid.gov.uk/News/Speeches-and-statements/2012/Justine-Greening-Eliminating-violence-against-women-and-girls/

Annex 4: DFID Theory of Change on Tackling Violence Against Women and Girls