Independent Quality Assessment of UNODC Evaluation Reports 2015

Synthesis Report

Prepared for:

Independent Evaluation Unit, UNODC

by John Mathiason and Ann Sutherland

November 22, 2016

CONTENTS

INTRODUCTION 3

METHODOLOGY 3

FINDINGS 4

A. Strengths of the Evaluation Reports 6

B. Weaknesses of the Evaluation Reports 6

C. Improvements from First Draft in Final 8

D. Best Practices Observed 9

D. Challenges for Evaluators 11

E. Mainstreaming Human Rights and Gender Considerations 11

RECOMMENDATIONS 12

Appendix 1. Comparison of Final Report with Draft Report 14

Appendix 2. UNODC EQA Template as Used in the Review 17

Appendix 3. Proposed Revised UNODC EQA Template 24

INTRODUCTION

The Independent Evaluation Unit (IEU) is leading and guiding evaluations in order to provide objective information on the performance of United Nations Office on Drugs and Crime (UNODC). As a member of the United Nations Evaluation Group (UNEG), IEU is following its Norms and Standards.

The number of independent project evaluations and in-depth evaluations has increased significantly over the past years at UNODC, generating important learning opportunities for all stakeholders, including IEU. UNODC's efforts to create better mechanisms for discussing and propelling strategic issues forward are being informed by IEU, whereby important insights are generated in the planning, implementation and finalization of evaluations.

Based on the pilot initiative in 2015 on the first Evaluation Quality Assessment (EQA) of all published evaluation reports from January 2014 to April 2015, IEU has decided to continue these efforts.

As a result, building on the first EQA in 2015 by further strengthening human rights and gender considerations in evaluations, this study continues to assess independently and objectively the quality of all published evaluation reports between 1 May 2015 and 15 December 2016 (2015: 9 in-depth and 6 project evaluations; an estimate for 2016 estimation of 3 in-depth and 19 project evaluations]. The present synthesized report covers 2015 including the results of the EQAs from 01 January to 30 April. Moreover, the quality of a random sample of 30% of first draft evaluation reports (per year) has been reviewed and assessed to analyze the qualitative changes between the first draft report and the published version.

This stage of the assignment took place from 15 October to 22 November 2016. It has been undertaken by two independent consultants - Dr. John Mathiason (Team Leader) and Ann Sutherland (Team Member). Both have extensive experience in conducting evaluations and meta-evaluations for international organizations. They are the Managing Director and Principal Associate, respectively, for Associates for International Management Services (AIMS).

METHODOLOGY

The first phase of this assignment involved making revisions, in collaboration with IEU, to the EQA template used last year. Key references in this review process were the updated UNEG Norms and Standards for Evaluation (June 2016) and the revised UNEG UN-SWAP Evaluation Performance Indicator (2014). The substantive changes thus made were:

- Adjusting some of the criteria categories. The three categories of (i) Background, (ii) Evaluation Purpose and Scope, and (iii) Evaluation Methodology were combined into two as follows: (1) Project/Programme Background and Evaluation Purpose, and (ii) Evaluation Scope and Methodology. A category of Reliability was added;
- Inclusion of additional criteria on human rights and gender;

- Addition of a new section to the EQA template that specifically addresses the extent to which each evaluation met the four key UN-SWAP indicators.

The reviewers examined the quality of all of the evaluation reports produced during 2015, including seven that had been reviewed for the previous synthesis. The total number of evaluations for the year was 22. Half of these were in-depth evaluation and half were independent evaluations. The reviewers also compared the final drafts of a third of the evaluations, randomly selected, with the first drafts to observe any differences that could be attributed to the process of commenting on the first drafts by IEU. Four of these were drafts of in-depth evaluations and three were drafts of independent evaluations.

The evaluation quality assessments used the new template for all of the evaluations that had not been previously assessed. To ensure consistency of the review process, both reviewers independently assessed three In-Depth and two Independent Project Evaluation reports. These reports were randomly selected and were rated according to the nine EQA criteria categories. The reviewers then compared their comments and scores for each criterion as well as the overall score. In all cases the overall scores were the same. There were minor differences in criteria scores and there were non-material differences in comments. The differences were discussed and resolved. As it was evident that there was consistency between the team, the remaining reports were each rated by one person; assignments were based on each reviewer's area of expertise and language skills. Team members consulted each other in the two cases where there was some uncertainty about the scoring. Overall, one third of the evaluation reports were assessed by both reviewers.

The team used the "Unsatisfactory" rating only when the criteria elements were missing or very poorly addressed. Similarly, "Very Good" was only used when all criteria were fully met. As a result, "Fair" was used when the criteria were generally not met, and therefore that score can be understood to mean that the reports, or pieces of the reports, were not well done. The Evaluation Quality Assurance (EQA) template used for the assessment is shown in the Annex.

FINDINGS

Overall, all of the 22 reports produced in 2015 were of adequate quality. As seen in Table 1, one report received a Very Good rating and no reports were rated as Unsatisfactory.

Table 1: Overall Rating

	Unsatisfactory	Fair	Good	Very Good	Total
# of Reports	0	12	9	1	22
% of Reports	0	55%	41%	5%	100%

There were, as last year, some differences by criterion. Table 2[1] shows that the criterion that was most likely to be fair or unsatisfactory was Scope and Methodology, largely because of problems in data collection including sampling. Content and Purpose was also relatively less strong, mostly because of inadequate description of the project and the frequent absence of logical frameworks. However, Lessons Learned and the overall Presentation and Structure are generally good. These results are slightly less favorable than those in 2014/15 but may reflect the more precise criteria in the new template.

Table 2: Report Rating by Criteria

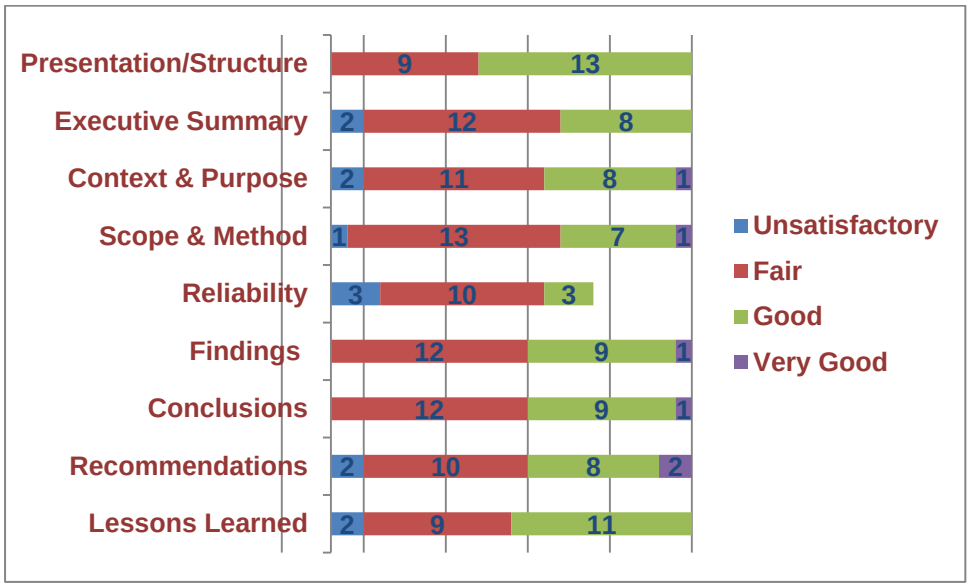
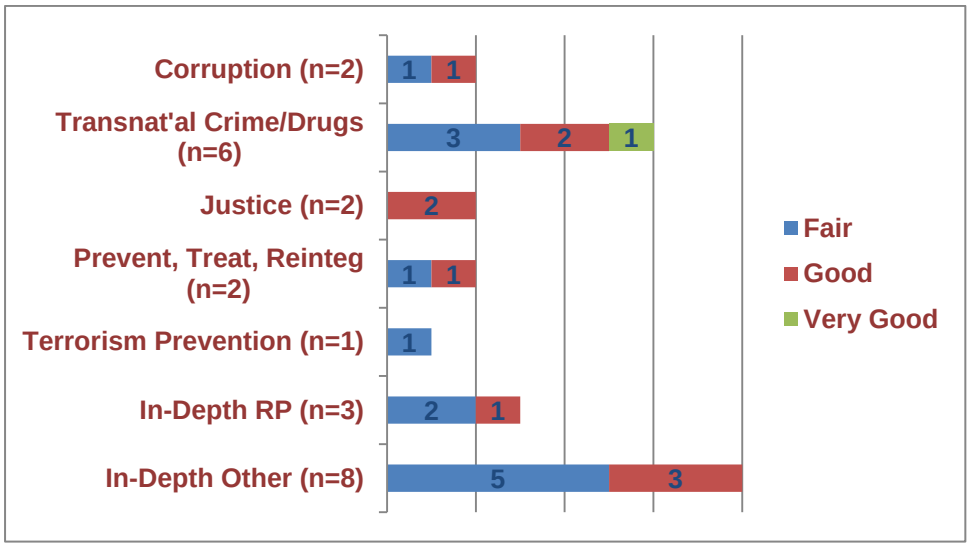


Table 3: Report Rating by Subject



[1] Note that fewer reports were rated for Reliability because this category was not included in the previous EQA review process.

A. Strengths of the Evaluation Reports

There were five primary strengths observed across the evaluation reports reviewed.

Consistency with DAC and UNEG Norms: The reports generally conformed to the accepted norms and guidelines.

Adherence to ToRs: For the most part, the reports adhered to their respective ToRs, within the limits of, sometimes, extensive ToR requirements (one ToR included 32 evaluation questions). In at least one case, the evaluators consolidated the questions so that the analysis was more readable.

Presentation and Structure: This section of the reports received the highest scores with 13 of 22 reports receiving a rating of Good. The most common concern is that some report use minimal visual aids or formatting techniques to break up extensive narrative and increase readability.

Scoring by Thematic Sector: The most highly rated theme was Justice – there were two reports within this theme and both were rated as Good. The only report to receive a rating of Very Good was one on Transnational Crime and Drugs. Overall, the Independent Project evaluations rated more highly than the In-Depth evaluations.

Human Rights and Gender: The extent to which human rights and gender are addressed within the projects/programmes under review is now consistently included in UNODC evaluation reports.

B. Weaknesses of the Evaluation Reports

There are eight general types of weaknesses that the reviewers identified.

Executive Summaries: This was one of two weakest sections of the report with only 36% of reports receiving a Good or higher rating. The primary concern is the length of this section. The IEU's guidelines reasonably state that the Executive Summary should be a maximum of 4 pages yet most reports exceed this length and some are as long as 11 pages. A common factor contributing to the excessive length is the inclusion of detailed Findings. Further weaknesses are the Purpose of the evaluation not being stated and the Methodology not being adequately specified. The description of the Methodology should be succinct yet sufficient to inform the reader of the robustness of the evaluation process.

Absence of Logframes: As was the case in the previous synthesis report, the most significant concern is the absence of results frameworks and clear indicators to measure the progress or achievement of the intervention against its planned results. The majority of the evaluation reports cited the lack of a strong logframe as a design or management flaw that prevented an adequate programme/project evaluation process from being carried out. Logframes were rarely included in the reports and only in one case did the evaluators prepare an improved framework (this was the one Very Good evaluation report). Revised Logframes should be mandatory components of evaluations since they are intended to express what has been promised in terms of outcomes and objectives to be achieved. If evaluators find existing ones inadequate, then there should

be some attempt to construct a logframe, or a least a basic results chain, in order to be clear about what the evaluation process is measuring. At a minimum, the evaluators should articulate the proposed chain of results or program theory.

Methodology Section: This continued to be a weak section of the reports with only 36% achieving a rating of good or higher. The major shortcomings are that:

- Most evaluators did not employ a sampling frame[2]:
- Those consulted during the evaluation process were not disaggregated by stakeholder group or other distinguishing factor such as gender, and in several cases, even just the total number of people consulted was not provided;
- Few evaluations included quantitative data collection processes even though this was often requested in the ToR. In the case of the KGZT90 report, a survey was conducted but no quantitative results were reported;
- Interview protocols (when provided) often replicate the key questions in the ToR. In such cases, the questions require an understanding of evaluation terminology that respondents are unlikely to have. An example is, "based on your experience, were the actions taken by <this project implementers> to achieve the project's outputs efficient?" (XSP/X70 report, p. 55). This report noted that respondents were frequently not able to answer the questions posed;
- The sources of data are often not evident in the findings or are not linked to the findings. This makes it difficult to know if the findings are just based on the perceptions of the evaluator;
- There was generally no indication of how data analysis processes were carried out.

Improvements noted in the Methodology section from the previous year are:

- Most reports include an annex with the list of stakeholders consulted. In most cases this did not include names or positions, but the organization and section of the organization represented are stated
- Reports consistently now include limitations of the study, although fewer indicate how the limitations were overcome.

Summary Matrix: There was inconsistency in how information is presented in the Evidence column. In most reports only the general source of information is cited (i.e. "documents, interviews"). Although there are instructions in the IEU Guidance for filling in the Findings and Recommendations' columns, there is no comparable explanation for Evidence. In fact, in one first draft of an evaluation, details were provided on the

[2] According to UNEG Standards, the sampling frame is expected to address: area and population to be represented, rationale for selection, mechanics of selection, numbers selected out of potential subjects, and limitations of the sample. Sampling is critical for ensuring that respondents are representative of key stakeholder groups, thereby ensuring the validity of the evaluation process.

evidence that showed clearly the basis for the finding, but that was converted into general sources in the final version of the document.

Reliability of Data: This was a new category added to reflect the importance of the quality of data and robustness of data collection processes. Only three of the 16 reports reviewed for this factor were rated as Good and three were Unsatisfactory. It is possible that the evaluators used more rigor in the their data collection approaches but did not document the processes in their reports.

Participation of Core Learning Partners (CLP)[3]: Although several reports described the main project stakeholders as having some level of involvement – primarily in the design of the evaluation (such as reviewing the Inception Report) – rarely were they noted as having a role in contributing to the conclusions or reviewing the draft report. In most cases, the involvement of the CLP appears to be restricted to providing input through regular data collection processes.

Use of Visual Aids: Visuals aids were used more frequently than in the previously reviewed reports, however in several cases there continued to be formatting mistakes that affected their usefulness in conveying information.

Scoring by Thematic Area: As noted, the In-Depth reports tended to be weaker overall than the Independent evaluations. The common issue of concern was the absence of an overall results framework, and this limited the quality of the Effectiveness section. This was particularly evident in the evaluation of Regional Programmes where the evaluators were challenged to assess the results of the overarching RP as well as, to varying extents, the sub-programmes and projects. In some cases the evaluators concluded it was not possible to assess the results of the RP itself and focused their efforts on the Design and Relevance sections.

C. Improvements from First Draft in Final

As requested, the reviewers compared the final drafts of seven evaluations with their first drafts on which IEU will have commented. These were randomly selected from the 22 evaluations. The results of the comparison are shown in Table 4 and the specific analysis is shown in Annex 1. Ultimately, there is no connection between changes made and the rating given to the evaluation.

As can be seen, the extent to which changes were made varies. In two cases, many changes were made, while in two no changes were made. For the most part, the re-drafted sections were more clearly formulated in one or more ways, i.e. additional evidence was provided, main findings were highlighted at the beginning or end of sub-sections, more 'diplomatic' language was inserted, and visual aids were added. However,

[3] Good practice in this regard can be found in the DAC guidelines which state that stakeholders should be given the opportunity to comment on findings, conclusions, recommendations and lessons learned. They further state that the evaluation report should reflect these comments and acknowledge any substantive disagreements.

not all of the changes were improvements – one draft report had a succinct Executive Summary that became densely written and overly lengthy in the final version. It is also likely that the final drafts do not include all annexes. An annex in one report included the results of an online survey - it takes up 60 pages because it was not formatted properly. This would generally be caught in a review process.

Table 4. Comparison of First and Final Drafts of a Sample of Evaluations

	Date of first / final draft	4+ months between versions	Substantive Changes	Final Rating
GMCP In-Depth - Global Maritime Crime (Justice)	Sept 2015 Oct 2015	No	Some	Fair
RP Eastern Africa Final	Jun 2015 Jun 2015	No	None	Fair
RP Arab States Final	Mar 2015 Jul 2015	Yes	Many	Fair
RP West Africa Final	Apr 2015 Dec 2015	Yes	Many	Good
BRAX63 Human Trafficking Brasil Final	Jan 2014 Jun 2015	Yes	None	Fair
RERE29 Precursor Control Central Asia and Azerbaijan Final	Nov 2014 Apr 2015	Yes	Some	Fair
KGZT90 Criminal Justice & Prison Reform Kyrgyz Republic Mid-term	May 2015 Aug 2015	No	Some	Good

Although assumptions can be made about the contribution of feedback from the IEU, it is not possible to attribute differences between the draft and final documents with any certainty as the text of the final reports did not note the type of input received.

There were differences in time between the first draft and the final. The average time was 5.4 months, ranging from almost no time (RP Eastern Africa Final) and almost a year and a half (BRAX63). In the latter case, one factor might be that the evaluation was written in Portuguese.

D. Best Practices Observed

Many examples of best practices were included in the previous summary report. These primarily focused on approaches that evaluators used to clearly specify the results and to increase the readability of the reports. Below are additional examples that mainly focus on improving evaluation quality.

Reconstruction of the logical framework:. A problem in several evaluations was that the evaluators found that the logical frameworks (or results matrices) did not indicate expected outcomes. In these cases, the evaluators took the logframes in the Terms of Reference as gospel and this contributed to the problem. In the evaluation rated very good (GLOU61 Global eLearning Program - Drugs, Crime & Terrorism), the evaluators, in the inception report, reconstructed the logical framework to show clearly expected outcomes and the expected causal connection with the project's output. The revised logframe that was used is shown in Exhibit 1.

Exhibit 1: Example of revised logical framework (excerpts)

Project Objective/ impact/ ultimate outcome: Enhancing law enforcement response to global human security challenges

Impact indicators: Effective identification, investigation and prosecution of transnational organized crimes, including trafficking in persons, drugs and precursors, natural resources and hazardous substances, and smuggling of migrants.

Outputs	Outcomes	Performance indicator/s	Means of Verification
Effective eLearning management and cost recovery models established and communicated within UNODC Global and regional eLearning workshops conducted Existing eLearning sites around the world assessed and upgraded Up-skill existing partners with current modules	**Outcome 1:** Law enforcement and other trained officials internalize and use new knowledge for protecting the rule of law and promoting health and development	Number of officers effectively informed / trained by country, organization, position, gender, and skill areas Evidence of improved operational performance back in the workplace	Pre and post test scores collected from training records Follow-up assessment studies/ survey
Support to member-states to make eLearning training programme compulsory for all national personnel assigned to law enforcement duties.	Member-states make training programme compulsory for all national personnel assigned to law enforcement duties.	Number of member-states that make training programme compulsory	Participating country agency training plans and records/reports and qualitative assessment by UNODC.
Support to member-states for institutionalizing CBT into ongoing basic law enforcement training including handing over equipment.	Member-states use and institutionalize CBT/ eLearning into ongoing basic law enforcement training programmes.	Number of member-states that institutionalize CBT into ongoing basic law enforcement training programmes.	Participating country agency training plans and records/reports and qualitative assessment by UNODC.
LMS produces timely and quality data on learning outcomes, student numbers and profiles, modules completed, etc. Periodic reviews of data quality and utility by CBTU in consultation with stakeholders	LMS data is effectively used by concerned stakeholders, including UNODC and participating government agencies/ training centre managers.	Number and type of stakeholders using LMS data	Participating country agency training records/reports and qualitative assessment by UNODC Feedback from stakeholders, using structured qualitative assessment tool

Regional Programme Evaluations: A good example of addressing the overall RP as well as its sub-programmes and projects is the evaluation of the Regional Programme for West Africa (2010-2014). Although the report has some problems, the formulation of the study and presentation of Findings could be a useful model for other RP assessments. The evaluators were able to effectively handle the 51 questions included in the ToR. An evaluation matrix with indicators was developed that showed how the main criteria and questions in the ToR would be addressed. In the Findings section, the main questions were highlighted in a box at the beginning of each criteria, and sub-headings were used to increase readability. In some but not all cases, summarizing statements were provided at the end of the subsection.

Linking Data Sources to Findings: The evaluation of GLOU61 Global eLearning Programme does a very good job of presenting the evidence on which the findings are based. The perspectives of stakeholders consulted are clearly shown, including visually through a word cloud. It is also one of the few reports where respondents were well disaggregated. Outside sources of data are also drawn upon and these strengthen the findings.

D. Challenges for Evaluators

There continue to be concerns about the ToRs that are provided to the evaluators. While these are very comprehensive, they may be too comprehensive (and as a result, too long). Some of the ToRs are up to 15 pages long. One reason they are so lengthy is that they indicate too many evaluation questions to be answered by the evaluation. Some ToRs included over 50 questions to be answered. This often means that the evaluators either ignore questions, respond with very general statements, or provide unnecessarily detailed findings. In contrast, one of the comparators used for this review, UNFPA, specifies a maximum of 8-10 questions, or one or two per category.

Many of the evaluations do not adequately distinguish between Effectiveness and Impact (the shorter and longer-term results). In some cases this distinction is not clarified in the ToRs.

E. Mainstreaming Human Rights and Gender Considerations

Human Rights and Gender are now consistently included in UNODC evaluation reports. Human rights are easier to include because the bulk of the projects being evaluated deal with issues of crime, where human rights are and always have been issues. Gender is a different case. This is addressed in regards to the programme or project being evaluated but are not considered in the methodology of the evaluation itself. At a minimum, it would be expected that the stakeholders consulted as part of the evaluation process be disaggregated by gender but this was rarely done. This was not done in the RER29 even though in its section on Human Rights and Gender it is stated that "it was however noteworthy, that a significant number of interviewees who were put forward by the beneficiary countries were female" (p. 17) and it points out that a weakness of the project is that it does not provided gender disaggregated data.

Part of the problem is that some of the projects being evaluated do not have an obvious gender dimension which in practice generally refers to women. An example is the project on maritime crime (or more precisely, piracy, mostly in Somalia) where none of the pirates are female. Of course, the gender dimension in that project deals with men. Determining what the gender dimension should be is a task of project formulation and should also be included in the Terms of Reference.

As far as the need to include gender in the methodology, it should not be sufficient to just state that gender was included as a cross-cutting issue in the methodology section without stating how the was accomplished (BGD/X79 is an example – p. 18).

There were several reports that had a strong analysis of Human Rights and Gender. These included: TIL/X78 Strengthening Land Border Control in Timor Leste, KGZT90 Criminal Justice and Prison Reform in Kyrgyz Republic, GMCP Global Maritime Crime, and RP for West Africa. The latter is notable for going beyond looking at numbers of women involved in various activities to addressing broader issues of equity and equality.

The following chart shows the average scores for the 15 reports assessed during this review process according to the UN-SWAP criteria. The SWAP tool assesses the extent to which gender equality and the empowerment of women (GEEW) is integrated into

evaluation processes. There are four criteria and each is rated on a scale of 0-3 with 0 being awarded when there is no integration, 1 when there is partial integration, 2 when there is satisfactory integration, and 3 when the evaluation process exceeds requirements.

The lowest score was for use of gender-responsive evaluation methodologies, and the highest was for including a gender analysis in the Findings, Conclusions, and Recommendations. The total average score for the evaluations reviewed was 5.5 which equates to an overall ranking of Fair.

Table 5: Average scores for the integration of GEEW (UN-SWAP)

Quality Assessment Criteria	Average Score (0-3)
a. GEEW is integrated in the evaluation scope of analysis and indicators are designed in a way that ensures GEEW-related data will be collected.	1.125
b. Evaluation criteria and evaluation questions specifically address how GEEW has been integrated into design, planning, implementation of the intervention and the results achieved.	1.625
c. Gender-responsive evaluation methodology, methods and tools, and data analysis techniques are selected.	.875
d. Evaluation findings, conclusions and recommendations reflect a gender analysis.	1.875
Overall Score	**5.5**
Overall Rating	**Fair**

RECOMMENDATIONS

The main recommendations for the IEU to consider, beyond those included in the previous EQA summary report, include:

Inclusion of results frameworks and/or the evaluation matrix: It is suggested that it be mandatory for at least one of these to be included as an annex to the reports.

Use of gender-responsive evaluation methodologies: Evaluators need to be made aware that they have to go beyond conducting an assessment of gender in the interventions they are studying, and use and report on gender-responsive processes within their own evaluation. The methodology needs to be explicit about how it is incorporating gender dimensions into the data gathering and assessment processes.

Noting changes made in the Inception Phase: The Evaluation Scope and Methodology Section should address whether the Inception Phase included substantive changes to processes outlined in the ToR, and if so, what these were.

Increased emphasis on results: Evaluations should focus more on the Effectiveness of the intervention and less on Design and Relevance, and possibly combine these sections. The Design and Relevance sections commonly comprise approximately half of the report Findings and, ultimately, all interventions are found to be relevant.

Use of formatting techniques to highlight key points: Lengthy narratives are less likely to be read by the intended audiences than more concise document where key points are highlighted. In addition to more use of visual aids, basic techniques to be encouraged include **use of bold font** for main themes or key points, **use of text boxes**, and **more numbering** of sections and sub-sections of the reports, as well as numbering the recommendations.

Attention to issues most often overlooked: There are a number of issues that are included in the UNODC criteria but have not generally been taken up and are now in the new UNEG Norms and Standards including: reports should have benchmarks for the criteria being assessed to the extent possible; the appropriateness of the evaluation team for the assignment should be noted; the reports should be explicit about stakeholder engagement; and they should be more being explicit about how gender and human rights are addressed in the evaluation process.

More guidance on support for in-depth evaluations: The need for additional support is most apparent for Regional Programmes, in particular in how the evaluators should address individual projects as part of the overall regional efforts. More guidance may be warranted for both how the ToRs are developed and for how IEU staff provide support throughout the evaluation process.

Further refinement of the EQA template: It is suggested that check boxes be incorporated into the template to make it easier to see the extent to which all of the main criterion have been addressed. The comments can then be less redundant (it will no longer be necessary to re-state the criteria) and can be used to highlight particular strengths and key areas of concern. A proposed format is attached.

Appendix 1. Comparison of Final Report with Draft Report

2015 BRAX63

The drafts were identical, with the exception that the first draft was submitted in January 2014 while the final was in June 2015. Also there was an English-language executive summary in the final and the summary matrix was moved up to after the Sumario Executivo.

GCMP

The final was in October 2015, the draft was September 2015. In the executive summary there was a major redraft of the beginning. A major redraft of the beginning which in the first draft read like a conclusion. Also, the summary now is structured according to the main question areas and is much longer. The original was about five pages, the final is 11.5 pages and was rated unsatisfactory. A new "major finding is that the GMCP remains relevant and meets the requirements of its stakeholders." Another finding now rated as key (previously only important) is "The rate of geographic and thematic expansion, well managed to date, has been relatively rapid. There is little evidence in project documentation that enough time has been dedicated to review and lesson learning." A new important recommendation: " The use of hard ear-marked funding reduces the flexibility and – on occasion – efficiency with which the GMCP operates. The funding modality of most UNODC projects and programmes relies on donors providing so called 'hard ear-marked' funding. This ties the funding to certain conditions and can hamper delivery of the overall objectives of the project or programme since the project or programme team does not have the latitude to utilize the funds in the way they deem best suited to the current situation. It can also lead to concerns that funding drives the strategy rather than strategy driving the funding as UNODC chases donor funds." Some graphs and text added. A new section on Quality of Design added. The text on effectiveness was doubled in the final. Recommendations were summarized for presentation, and became more specific.

RER/E29

Draft was November 2014, Final was April 2015. Considerable changes in the executive summary which made it slightly shorter but still too long. Changes in the summary matrix with addition of some findings. The data in the matrix on evidence was more complete in the draft and in the final was made much shorter. More material added in the recommendations on Partnerships. A section on human rights and gender was added. Recommendations were increased (double the size of the first draft). Also Lessons Learned was doubled in size.

RP East Africa

Both reports are dated June 2015. There are no substantive differences between the two versions of the report. The only changes in the final report were: page headers were added; a pie chart from the ToR was copied into the main body of the report; an Annex for List of Projects was removed (the draft only had the heading and not the contents). There were minor editing problems, such as the mis-numbering of one of the annexes in the body of the report, that were not caught in the review process.

RP West Africa

Draft was April 2015, Final was December 2015. There are substantive differences between the two documents that improved the final version. The Executive Summary was not included in the draft. The Background section included more footnoted citations for sources of data, however, several of these included rough notes presumably from the reviewer that questioned the completeness of the citation.

The final version of the report has been heavily edited - the amount of text (detail) has been reduced throughout and the wording has been sharpened. For the most part, this has made the report clearer. Some of the new phrasing has softened the tone of the report with shortcomings being phrased less critically. A striking example is in the Efficiency section where the draft concludes that efficiency has "somewhat improved over time" and the qualifier of "somewhat" is removed in the final report (p. 48).

In a few cases the detail that was removed included supporting information, i.e., the Methodology in the draft showed somewhat more rigor as it had useful information about the purpose of data collection tools for different stakeholder groups and noted that a local evaluator from Sierra Leone was engaged as a junior consultant to address access constraints faced by the team. The final version was improved by adding the percent of respondents to the online questionnaire.

The Findings section shows the most improvement including through the addition of summary statements of findings in some of the longer subsections in Effectiveness. However, this was not done in a consistent manner – some these statements appeared at the beginning and sometimes at the end, and in one case bold typeface was used. The Impact section was substantially redone to focus more on longer term impacts; the draft focused more on short-term results and design issues.

A section that was not improved is Recommendations – the number of recommendations was increased (from 18 to 23, although the actual number is higher in both documents as several had sub-recommendations) and the lack of clarity was not addressed. On the other hand, the Lessons Learned section was better formulated and the number of lessons was reduced. One annex includes the results of the online survey - it takes up 60 pages because it was not formatted properly and it is questionable whether it needed to be included at all as the response rate was so low.

RP Arab States

The draft report was produced in March and the final in July 2015. The changes in the two documents are fairly substantial. Although there was some streamlining of text, the final document is longer in several sections as more explanation / supporting evidence has been included. The additional detail is mostly helpful (i.e., adding a major achievement of the Palestine program, a fuller explanation of Full Cost Recovery which would be helpful to some readers as the concept was found to not be well understood); a notable exception is the Executive Summary which expanded from 3 pages to 5 pages with more detailed findings than necessary.

The final version is heavily footnoted, including in the Executive Summary, adding to the overall length. Editing and formatting increased the overall clarity of the report. Bolding was effectively used to highlight the topics in the Summary Matrix, Visual aids were somewhat better explained although consistent formatting was not used and did not adhere to guidelines for titles.

Substantive improvements include:

- the addition of methodology in the Executive Summary
- better grouping and somewhat clarifying the recommendations in the summary matrix
- adding, or improving, statements at the end of sub-sections to highlight findings
- building on a weak Effectiveness section by including accomplishments and challenges within each sub-program (although effectiveness of the overall program is not addressed)
- building on a weak Impact section although it was overly focused on short-term results that would have been better placed in Effectiveness.
- expanding and improving the gender section to include more of a GEEW perspective
- making the recommendations more clear and grouping them, although the descriptions for some were too lengthy.

KGZT90

The draft was produced in May and the final in August 2015. There have been some changes and additions that have strengthened the report. The final document is easier to read – long paragraphs have been broken up and some of the text has been sharpened and/or made more succinct. Substantive changes include increased reference to sources of data, more information in Efficiency, and the inclusion of the summary matrix, a table linking project activities to UNODC program areas, and the project results framework in Effectiveness. The order of the recommendations was shifted to reflect their importance. One recommendation was added to the already large number (22) and it is about competitive tenders, an issue that was not addressed earlier in the report.

Appendix 2. UNODC EQA Template as Used in the Review

UNODC EQA Template

General Project Information	
Project/Programme Number and Name	
Thematic Area	
Geographic Area (Region, Country)	
Approved budget at time of the evaluation (USD)	
Type of Evaluation (In-Depth/Independent Project; final/ midterm; other)	
Evaluator(s)	
Date of Evaluation (from MM/YYYY to MM/YYYY)	
Date of Evaluation Report (MM/YYYY)	
Quality Assessment conducted on/by	

OVERALL QUALITY RATING:

SUMMARY:

Quality Assessment Criteria	Assessment Levels			
	Very Good	Good	Fair	Unsatisfactory
1. Structure, Completeness And Clarity Of Report a. Format (headings, font) accords to IEU Guidelines and Templates for Evaluation Reports. b. Structure accords to IEU Guidelines for Evaluation Reports with the following sequence: Executive Summary; Summary Matrix of Findings, Evidence and Recommendations; Introduction (Background and Context, Evaluation Scope and Methodology, Limitations to the Evaluation); Findings (Design, Relevance, Efficiency, Partnership and Cooperation, Effectiveness, Impact, Sustainability, Human Rights and Gender Equality/mainstreaming, Innovation); Conclusions; Recommendations; Lessons Learned. c. Language is empowering and inclusive avoiding gender, heterosexual, age, cultural and religious bias, among others. d. Report is easy to read and understand (i.e. written in an accessible non-technical language appropriate for the intended audience). Visual aids, such as maps and graphs, are used to convey key information. List of acronyms is included. e. Report is free from any grammar, spelling, or punctuation errors. f. Objectives stated in the terms of reference are adequately addressed. g. Issues of human rights are adequately addressed h. Issues of gender equality/mainstreaming are adequately addressed. i. Report contains a logical sequence: evidence-assessment-findings-conclusions-recommendations. j. Composition of Evaluation Team is included and has gender and geographic expertise. Preferably it is gender balanced and includes professionals from countries or regions concerned. k. Annexes include at a minimum: evaluation terms of reference; list of persons interviewed and sites visited; list of documents consulted; evaluation tools used.				
2. Executive Summary a. Written as a stand-alone section that provides an overview of the				

evaluation and presents its main results. b. Generally follows the structure of: i) Purpose, including intended audience(s); ii) Objectives and brief description of intervention; iii) Methodology); iv) Main Conclusions; v) Recommendations. c. Summary Matrix presents only the key and most important recommendations from evaluation report. d. Findings, sources and recommendations in the Summary Matrix are clear and cohesive, and specify the stakeholder to whom they are addressed. e. Maximum length 4 pages, excluding the Summary Matrix.	.
3. Evaluation Context And Purpose a. Clear description of the project evaluated is presented. b. Logic model and/or the expected results chain, and /or program theory (that at a minimum identifies and links objectives, outcomes and indicators of the project) is clearly described. c. Context of key cultural, gender related, social, political, economic, demographic, and institutional factors are described, and the key stakeholders involved in the project implementation and their roles are identified. d. Project's status is described including its phase of implementation and any significant changes (e.g. strategies, logical frameworks) that have occurred. e. Purpose of the evaluation is clearly defined, including why it was needed at that point in time, who needed the information, what information is needed, how the information will be used, and the target audience.	.
4. Scope And Methodology a. Evaluation scope is clearly explained including the main evaluation criteria, questions and justification of what the evaluation did and did not cover. b. Transparent description presented of methodology applied; how it was designed to address the evaluation purpose, objectives, questions and criteria is explained. c. Methodology allows for drawing causal connections between output and expected outcomes d. Gender sensitive methodology aware of power relations during an evaluation process, inclusive and participatory.	

e. Data collection methods and analysis, and data sources are clearly described; as are the rationale for selecting them, and their limitations are clearly described. Reference indicators and benchmarks are included where relevant. f. Sampling frame clearly described and includes area and population to be represented, rationale for selection, mechanics of selection including whether random, numbers selected out of potential subjects, and limitations of sample. g. Methods are appropriate for analysing gender equality/mainstreaming and human rights issues identified in evaluation scope h. High degree of participation of internal and external stakeholders, including the Core Learning Partners, throughout the evaluation process is planned for and made explicit. When there are thematic or approach gaps (i.e. gender equality/mainstreaming) among stakeholders, other key informants not directly involved in the project were invited for consultation.	
5. Reliability of Data *To ensure quality of data and robust data collection processes* a. Triangulation principles (using multiple sources of data and methods) were applied to validate findings. b. Qualitative and quantitative data sources were used, and included the range of stakeholder groups and additional key informants (when necessary) defined in evaluation scope. c. Limitations that emerged in primary and secondary data sources and collection processes (bias, data gaps, etc.) are identified and, if relevant, actions taken to minimize such issues are explained. d. Evidence provided of how data was collected with a sensitivity to issues of discrimination and other ethical considerations. e. Adequate disaggregation of data by relevant stakeholder undertaken (gender, ethnicity, age, under-represented groups, etc.). If this has not been possible, it is explained.	
6. FINDINGS AND ANALYSIS To ensure sound analysis and credible findings *Findings*	

a. Have been formulated clearly, take into account any identified benchmarks, and are based on rigorous analysis of the data collected. b. Address all evaluation criteria and questions raised in the ToR including relevance, efficiency, effectiveness, impact and sustainability, as well as UNODC's additional criteria of design, partnership and cooperation, innovation, and the cross-cutting themes of human rights and gender. c. Address any limitations or gaps in the evidence and discuss any impacts on responding to evaluation questions raised in ToR. d. Discuss any variances between planned and actual results of the project (in terms of objectives, outcomes, outputs). e. Are presented in a clear manner. *Analysis* a. Interpretations are based on carefully described assumptions. b. Contextual factors are identified (including reasons for accomplishments and failures, and continuing constraints). c. Cause and effect links between an intervention and its end results (including unintended results) are explained. d. Includes substantive analysis of gender equality/mainstreaming issues e. Includes substantive analysis of human rights issues.	
7. CONCLUSIONS a. Take into consideration all evaluation criteria and questions, including human rights and gender equality/mainstreaming criteria. b. Have been formulated clearly and are based on findings and substantiated by evidence collected. c. Convey the evaluators' unbiased judgement of the intervention d. Developed with the involvement of relevant stakeholders. e. Present a comprehensive picture of both the strengths and weaknesses of the project. f. Go beyond the findings and provide a thorough understanding of the underlying issues of the project and add value to the findings.	
8. RECOMMENDATIONS a. Are clearly formulated, based on the conclusions, and substantiated by evidence collected.	

b. Address flaws, if any, in project's data acquisition processes. c. Are specific, realistic, time-bound and actionable, and of a manageable number. d. Are clustered and prioritized. e. Reflect stakeholders' consultations whilst remaining balanced and impartial f. Clearly identify a target group for action.	
9. LESSONS LEARNED a. Are correctly identified, innovative and add value to common knowledge. b. Are based on specific evidence and analysis drawn from the evaluation. c. Have wider applicability and relevance to the specific subject and context.	.

SCORING

Element Of The Evaluation	Points Per Category	Points Awarded			
		Very Good	Good	Fair	Unsatisfactory
Presentation And Completeness	10				
Executive Summary	5				
Evaluation Context And Purpose	5				
Evaluation Scope And Methodology	10				
Reliability Of Data	5				
Findings And Analysis	35				
Conclusions	10				
Recommendations	15				
Lessons Learned	5				
Total Maximum Score	**100**				
		Very Good -> very confident to use	**Good ->** confident to use	**Fair ->** use with caution	**Unsatisfactory ->** not confident to use

ASSESSMENT OF THE INTEGRATION OF GENDER EQUALITY AND EMPOWERMENT OF WOMEN (GEEW) for UN-SWAP

Quality Assessment Criteria	Comments	Score (0-3)
e. GEEW is integrated in the evaluation scope of analysis and indicators are designed in a way that ensures GEEW-related data will be collected.		
f. Evaluation criteria and evaluation questions specifically address how GEEW has been integrated into design, planning, implementation of the intervention and the results achieved.		
g. Gender-responsive evaluation methodology, methods and tools, and data analysis techniques are selected.		
h. Evaluation findings, conclusions and recommendations reflect a gender analysis.		
Overall Score		
Overall Rating		

UN-SWAP Scoring System

Exceeding Requirements	3 - Fully integrated. Applies when all of the elements under a criterion are met, used and fully integrated in the evaluation and no remedial action is required
Meeting Requirements	2 - Satisfactorily integrated. Applies when a satisfactory level has been reached and many of the elements are met but still improvement could be done
Approaching Requirements	1 - Partially integrated. Applies when some minimal elements are met but further progress is needed and remedial action to meet the standard is required.
Missing	0 - Not at all integrated. Applies when none of the elements under a criterion are met.
Overall Calculation	11-12 = very good 8-10 = good 4-7 = Fair 0-3=unsatisfactory

Appendix 3. Proposed Revised EQA Template

UNODC EQA Template

General Project Information	
Project/Programme Number and Name	
Thematic Area	
Geographic Area (Region, Country)	
Approved budget of the time of the evaluation (USD)	
Type of Evaluation (In-Depth/Independent Project; final/ midterm; other)	
Evaluator(s)	
Date of Evaluation (from MM/YYYY to MM/YYYY)	
Date of Evaluation Report (MM/YYYY)	
Quality Assessment conducted on/by	

OVERALL QUALITY RATING:

SUMMARY:

Quality Assessment Criteria	Assessment Levels: Very Good - Good - Fair - Unsatisfactory Meets Criteria: Y = Yes N = No P = Partially	
1. Structure, Completeness And Clarity Of Report	**RATING:**	
l. Structure accords to IEU Guidelines for Evaluation Reports with the following sequence: Executive Summary; Summary Matrix of Findings, Evidence and Recommendations; Introduction (Background and Context, Evaluation Scope and Methodology, Limitations to the Evaluation); Findings (Design, Relevance, Efficiency, Partnership and Cooperation, Effectiveness, Impact, Sustainability, Human Rights and Gender Equality/mainstreaming, Innovation); Conclusions; Recommendations; Lessons Learned.		
m. Language is empowering and inclusive avoiding gender, heterosexual, age, cultural and religious bias, among others.		
n. Report is easy to read and understand (i.e. written in an accessible non-technical language appropriate for the intended audience). Visual aids, such as maps and graphs, are used to convey key information. List of acronyms is included.		
o. Report is free from any grammar, spelling, or punctuation errors.		
p. Objectives stated in the terms of reference are adequately addressed.		
q. Issues of human rights and gender equality/mainstreaming are adequately addressed		
r. Report contains a logical sequence: evidence-assessment-findings-conclusions-recommendations.		
s. Composition of Evaluation Team is included and has gender and geographic expertise. Preferably it is gender balanced and includes professionals from countries or regions concerned.		
t. Annexes include at a minimum: evaluation terms of reference; list of persons interviewed and sites visited; list of documents consulted; evaluation tools used.		
2. Executive Summary	**RATING:**	
f. Written as a stand-alone section that provides an overview of the evaluation and presents its main results.		
g. Generally follows the structure of: i) Purpose, including intended audience(s); ii) Objectives and brief description of intervention; iii) Methodology); iv) Main Conclusions; v) Recommendations.		
h. Summary Matrix presents only the key and most important recommendations from evaluation report.		

i. Findings, sources and recommendations in the Summary Matrix are clear and cohesive, and specify the stakeholder to whom they are addressed.		
j. Maximum length 4 pages, excluding the Summary Matrix.		
3. Evaluation Context And Purpose	**RATING:**	
a. Clear description of the project evaluated is presented.		
b. Logic model and/or the expected results chain, and /or program theory (that at a minimum identifies and links objectives, outcomes and indicators of the project) is clearly described		
c. Context of key cultural, gender related, social, political, economic, demographic, and institutional factors are described, and the key stakeholders involved in the project implementation and their roles are identified.		
d. Project's status is described including its phase of implementation and any significant changes (e.g. strategies, logical frameworks) that have occurred.		
e. Purpose of the evaluation is clearly defined, including why it was needed at that point in time, who needed the information, what information is needed, how the information will be used, and the target audience.		
4. Scope And Methodology	**RATING:**	
a. Evaluation scope is clearly explained including the main evaluation criteria, questions and justification of what the evaluation did and did not cover.		
b. Transparent description presented of methodology applied; how it was designed to address the evaluation purpose, objectives, questions and criteria is explained.		
c. Methodology allows for drawing causal connections between output and expected outcomes		
d. Gender sensitive methodology aware of power relations during an evaluation process, inclusive and participatory.		
e. Data collection methods and analysis, and data sources are clearly described; as are the rationale for selecting them, and their limitations are clearly described. Reference indicators and benchmarks are included where relevant.		
f. Sampling frame clearly described and includes area and population to be represented, rationale for selection, mechanics of selection including whether random, numbers selected out of potential subjects, and limitations of sample.		

g. Methods are appropriate for analysing gender equality/mainstreaming and human rights issues identified in evaluation scope		
h. High degree of participation of internal and external stakeholders, including the Core Learning Partners, throughout the evaluation process is planned for and made explicit. When there are thematic or approach gaps (i.e. gender equality/mainstreaming) among stakeholders, other key informants not directly involved in the project were invited for consultation.		
5. Reliability of Data *To ensure quality of data and robust data collection processes*	**RATING:**	
a. Triangulation principles (using multiple sources of data and methods) were applied to validate findings.		
b. Qualitative and quantitative data sources were used, and included the range of stakeholder groups and additional key informants (when necessary) defined in evaluation scope.		
c. Limitations that emerged in primary and secondary data sources and collection processes (bias, data gaps, etc.) are identified and, if relevant, actions taken to minimize such issues are explained.		
d. Evidence provided of how data was collected with a sensitivity to issues of discrimination and other ethical considerations.		
e. Adequate disaggregation of data by relevant stakeholder undertaken (gender, ethnicity, age, under-represented groups, etc.). If this has not been possible, it is explained.		
6. FINDINGS AND ANALYSIS *To ensure sound analysis and credible findings*	**RATING:**	
Findings	-	
f. Have been formulated clearly, take into account any identified benchmarks, and are based on rigorous analysis of the data collected.		
g. Address all evaluation criteria and questions raised in the ToR including relevance, efficiency, effectiveness, impact and sustainability, as well as UNODC's additional criteria of design, partnership and cooperation, innovation, and the cross-cutting themes of human rights and gender.		
h. Address any limitations or gaps in the evidence and discuss any impacts on responding to evaluation questions raised in ToR.		

i. Discuss any variances between planned and actual results of the project (in terms of objectives, outcomes, outputs).		
j. Are presented in a clear manner.		
Analysis	-	
f. Interpretations are based on carefully described assumptions.		
g. Contextual factors are identified (including reasons for accomplishments and failures, and continuing constraints).		
h. Cause and effect links between an intervention and its end results (including unintended results) are explained.		
i. Includes substantive analysis of gender equality/mainstreaming issues		
j. Includes substantive analysis of human rights issues.		
7. CONCLUSIONS	**RATING:**	
g. Take into consideration all evaluation criteria and questions, including human rights and gender equality/mainstreaming criteria.		
h. Have been formulated clearly and are based on findings and substantiated by evidence collected.		
i. Convey the evaluators' unbiased judgement of the intervention		
j. Developed with the involvement of relevant stakeholders.		
k. Present a comprehensive picture of both the strengths and weaknesses of the project.		
l. Go beyond the findings and provide a thorough understanding of the underlying issues of the project and add value to the findings.		
8. RECOMMENDATIONS	**RATING:**	
g. Clearly formulated, based on the conclusions, substantiated by evidence collected.		
h. Address flaws, if any, in project's data acquisition processes.		
i. Are specific, realistic, time-bound and actionable, and of a manageable number.		
j. Are clustered and prioritized.		
k. Reflect stakeholders' consultations whilst remaining balanced and impartial		
l. Clearly identify a target group for action.		
9. LESSONS LEARNED	**RATING:**	
d. Are correctly identified, innovative and add value to common knowledge.		
e. Are based on specific evidence and analysis drawn from the evaluation.		
f. Have wider applicability and relevance to the specific subject and context.		

SCORING

Element Of The Evaluation	Points Per Category	Points Awarded			
		Very Good	Good	Fair	Unsatisfactory
Presentation And Completeness	10				
Executive Summary	5				
Evaluation Context And Purpose	5				
Evaluation Scope And Methodology	10				
Reliability Of Data	5				
Findings And Analysis	35				
Conclusions	10				
Recommendations	15				
Lessons Learned	5				
Total Maximum Score	**100**				
		Very Good -> very confident to use	**Good ->** confident to use	**Fair ->** use with caution	**Unsatisfactory ->** not confident to use

ASSESSMENT OF THE INTEGRATION OF GENDER EQUALITY AND EMPOWERMENT OF WOMEN (GEEW) for UN-SWAP

Quality Assessment Criteria	Comments	Score (0-3)
i. GEEW is integrated in the evaluation scope of analysis and indicators are designed in a way that ensures GEEW-related data will be collected.		
j. Evaluation criteria and evaluation questions specifically address how GEEW has been integrated into design, planning, implementation of the intervention and the results achieved.		
k. Gender-responsive evaluation methodology, methods and tools, and data analysis techniques are selected.		
l. Evaluation findings, conclusions and recommendations reflect a gender analysis.		

Overall Score		
Overall Rating		

UN-SWAP Scoring System

Exceeding Requirements	3 - Fully integrated. Applies when all of the elements under a criterion are met, used and fully integrated in the evaluation and no remedial action is required
Meeting Requirements	2 - Satisfactorily integrated. Applies when a satisfactory level has been reached and many of the elements are met but still improvement could be done
Approaching Requirements	1 - Partially integrated. Applies when some minimal elements are met but further progress is needed and remedial action to meet the standard is required.
Missing	0 - Not at all integrated. Applies when none of the elements under a criterion are met.
Overall Calculation	11-12 = very good 8-10 = good 4-7 = Fair 0-3=unsatisfactory