# Third party Quality Check for Ex-post Evaluation (Internal Evaluation) in JICA

# Analysis Report in 2019

Consultant: KPMG AZSA LLC

# Overview of the targeted projects of third party QC in 2019

- Targeted projects: 50 (including 2 integrated evaluations)
- Scheme: Technical cooperation projects (TC): 34 (including 1 integrated evaluation), technical cooperation for development planning (DP): 4, grant aid (GA): 12
- Overall rating: highly satisfactory: 17 (including 1 integrated evaluation), satisfactory: 20, partially satisfactory: 9, unsatisfactory: 4
  Because the population of targeted projects (projects for which ex-post evaluation was conducted in 2017) is as small as 100 or less, the selection of target projects is not random sampling, but is based on the balance of sectors and schemes. Projects that seems to be necessary to be reviewed were Intentionally extracted. (This report describes the changes between the previous year (2018) and the current year (2019), but it is necessary to keep in mind that how the target projects are selected is different and cannot be simply compared.)

## List of the targeted projects

| No. | Country | Name of the project | Scheme |
|---|---|---|---|
| 1 | Indonesia | The Project for Research and Education Development on Information and Communication Technology in Institut Teknologi Sepuluh Nopember, (Phase 1 and Phase 2) | TC |
| 2 | Indonesia | Identification of Anti-Hepatitis C Virus (HCV) Substances and Development of HCV and Dengue Vaccines (SATREPS) | TC |
| 3 | Cambodia | The Project for Capacity Development for Implementing the Organic Law at Capital & Provincial Level | TC |
| 4 | Laos | Project on Riverbank Protection Works Phase II | TC |
| 5 | Laos | The Project For Introduction Of Clean Energy By Solar Electricity Generation System | GA |
| 6 | Viet Nam | Electric Power Technical Standards Promotion in Vietnam | TC |
| 7 | Viet Nam | Strengthening the Capacities for the Field of Management of Vietnam's Crop Production Sector for Improving the Productivity and Quality of Crop's Products | TC |
| 8 | Viet Nam | Project for Implementing Maternal and Child Health Handbook for Scaling Up Nationwide | TC |
| 9 | Viet Nam | Project for Strengthening Capacity of Inspection System for Ensuring Safety of Agro-Fishery Foods | TC |
| 10 | Myanmar | Strengthening Capacity of Training Teams for Basic Health Staff | TC |
| 11 | Myanmar | The Project For Provision Of Road Construction And Maintenance Equipment In Kayin State | GA |
| 12 | Philippines | National Industry Cluster Capacity Enhancement Project | TC |
| 13 | Tonga | The Project For Introduction Of Clean Energy By Solar Home System | GA |
| 14 | Vanuatu | The Project for Promotion of the Grace of the Sea in Coastal Village in Vanuatu (Phase 1 and Phase 2) | TC |

# List of the targeted projects

| No. | Country | Name of the project | Scheme |
|---|---|---|---|
| 15 | Papua New Guinea | The Project For Improvement Of Road Maintenance Equipment | GA |
| 16 | China | Project for Capacity Development on Mental Health Services for Reconstruction Support of Sichuan Earthquake | TC |
| 17 | China | Project on forest restoration after the earthquake in Sichuan Province | TC |
| 18 | Azerbaijan | The Project For Improvement Of Equipment For Amelioration And Irrigation (Phase 2) | GA |
| 19 | Afghanistan, Tajikistan | Rural Development Project in Tajik-Afghan Border Area of Gorno-Badakhshan Autonomous Oblast | TC |
| 20 | Bhutan | Study on GLOFs (Glacial Lake Outburst Floods) in the Bhutan Himalayas (SATREPS) | TC |
| 21 | Pakistan | THE DISTRICT HEALTH INFORMATION SYSTEM PROJECT FOR EVIDENCE-BASED DECISION MAKING AND MANAGEMENT | TC |
| 22 | Sri Lanka | The Project for Construction of a Dredger | GA |
| 23 | Nepal | Participatory Watershed Management and Local Governance Project | TC |
| 24 | Guatemala | Project for Child Health in Department of Quetzaltenango | TC |
| 25 | Guatemala | Project for Maternal and Child Health in Quetzaltenango,Totonicapan, and Solola in the Republic of Guatemala | TC |
| 26 | Dominican Republic | Sustainable Tourism based on Public-Private Partnership | TC |
| 27 | Honduras | The Project For Restoration Of The Democracia Bridge | GA |
| 28 | Mexico | Project for Human Resource Development in the technology of Plastic Transformation | TC |
| 29 | Bolivia | Project for strengthening health network in Rural region focusing on mother and children health | TC |
| 30 | Kenya | Project on Integrated Urban Development Master Plan for the City of Nairobi | DP |
| 31 | Uganda | Secondary Science and Mathematics Teacher's Project | TC |
| 32 | Uganda | District and Urban Roads (DUR) Mapping and Roads Database Project | TC |
| 33 | Ethiopia | Strengthening Infectious Disease Prevention, Control and Response in Amhara Region | TC |
| 34 | Ethiopia | National Pilot Project for Strengthening Mathematics and Science Education | TC |

# List of the targeted projects

| No. | Country | Name of the project | Scheme |
|---|---|---|---|
| 35 | Ghana | The Project For Introduction Of Clean Energy By Solar Electricity Generation System | GA |
| 36 | Sierra Leone | Integrated Project for Rural Health Improvement | TC |
| 37 | Comoros | Project for Capacity Development of the National School of Fisheries | TC |
| 38 | Cameroon | The Project For Rural Water Supply (Phase V) | GA |
| 39 | The former Yugoslav Republic of Macedonia | Project on Development of Integrated System for Prevention and Early Warning of Forest Fires | TC |
| 40 | Thailand | Project for flood countermeasures for Thailand agricultural sector | DP |
| 41 | Myanmar | The Project for Construction of Primary School-cum-Cyclone Shelter in the Area Affected by Cyclone "Nargis" | GA |
| 42 | Morocco | The Project for Flood Forecasting and Warning System in High Atlas Area | GA |
| 43 | Ethiopia | Project for Improving Maternal and Child Nutrition status | TC |
| 44 | Zambia | The project for Strengthening Community-based Child Health Promotion System in Urban Areas | TC |
| 45 | Zimbabwe | Project for the Improvement of Water Supply, Sewage and Solid Waste Management in Chitungwiza | DP |
| 46 | Mozambique | The Project for the Comprehensive Urban Transport Master Plan for the Greater Maputo | DP |
| 47 | Lesotho | The Project for Introduction of Clean Energy by Solar Electricity Generation System | GA |
| 48 | Niger | Support to the improvement of school management through Community Participation in Niger (School for all) Phase 2 | TC |
| 49 | Burkina Faso | School Management Committee Support Project | TC |
| 50 | Turkey | School-based Disaster Education Project | TC |

Note: SATREPS means "Science and Technology Research Partnership for Sustainable Development".

# Objectives and background

JICA evaluates projects for mainly fulfilling two purposes: 1) improve the projects further through the PDCA cycle and 2) ensure accountability to stakeholders including Japanese nationals and people of the partner. Internal ex-post evaluation is conducted by JICA overseas office staff and regional departments in the headquarter. The targets of internal ex-post evaluation are the projects which cost 200 million yen or more and under one billion yen.

In order to achieve the objectives of evaluation more effectively and efficiently, the internal evaluation process is being improved by introducing self-assessment by internal evaluators and quality check by external third parties.

The objectives of this study is to examine internal evaluations by external third party and to ensure the quality, objectivity, and impartiality of future internal evaluations as well as to disclose its outline to the public to strengthen the accountability.

# Third party QC tools and information collection methods

1. Third party QC is conducted through the desktop research. External third parties examine internal evaluation report and verify them based on a check sheet similar to the one developed for self-assessment.

2. The person in charge of third party QC ticks Yes/ Partly Yes/ No/ Not Applicable for each checklist item and comment the reasons on the check sheet.

3. Overall trends (e.g. average and characteristic results etc.) are analyzed based on the result of third party QC of individual projects.

4. For analysis of the overall trends, the results of the third party QC are calculated as score.

5. Information source for third party QC is limited to the internal evaluation report in principle.

6. Third party QC check sheet is aligned with the JICA Guidelines for Project Evaluation and Internal Evaluation Implementation Manual.

7. This year is the second third party QC following 2018. After the third party QC in 2018, based on that experience, the check sheet was partially revised and the scoring standards were clarified. (This report describes the changes over the past year (2018) and this year (2019). However, it should be noted that simple comparison are not always possible due to slight revisions to the scoring standards.)

# Methodology of scoring

① **Raw score** :
   Score 2 points for "Yes", 1 point for "Partly Yes", 0 point for "No", and no score for "Not Applicable".

② **Standardized score** :
   Standardized score = total of raw score ÷ (total number of checklist items − number of tick on "Not Applicable" )×2
   Standardized score is between 0.0 to 1.0. As the score is closer 1, it means that there are more ticks on "Yes", in other words, the internal evaluation's quality is higher.

[Note : Calculation of dispersion and standard deviation]

$$\text{Dispersion :} \quad \frac{1}{n}\sum_{n=1}^{n}(x_i - \bar{x})^2 \qquad \text{Standard deviation :} \quad \sigma = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(x_i - \bar{x})^2}$$

$n$ = total number of targeted projects、$x_i$ = average of standardized score、$\bar{x}$ = standardized score

The dispersion and standard deviation were calculated for the whole and each evaluation criteria, and the variation of the third party QC results of the target projects was confirmed. The greater the dispersion and standard deviation, the greater the variation from the average point.

# Analysis in this report

1. Overall trends

2. Evaluation criteria（①Relevance, ②Effectiveness / Impact, ③Efficiency, ④Sustainability, ⑤Conclusion, Recommendations and Lessons Learned, ⑥General Matters）

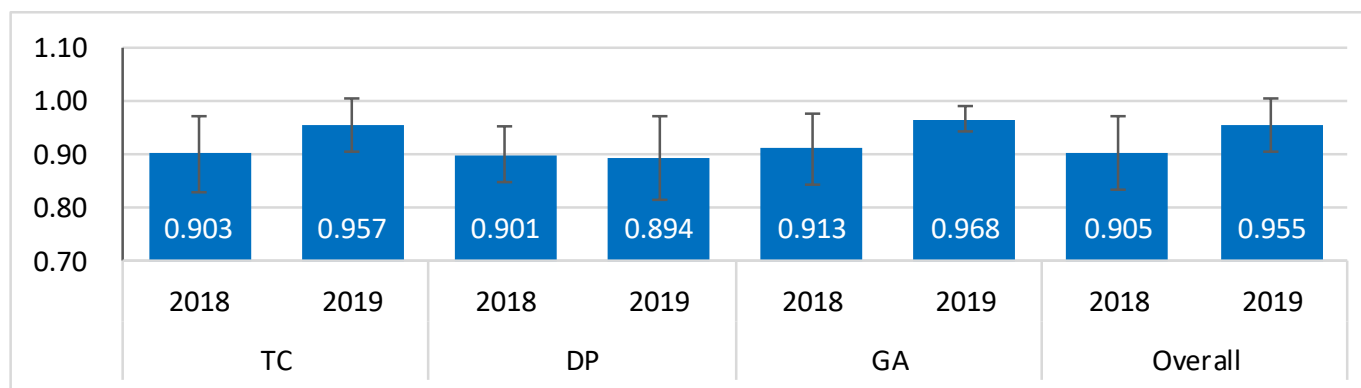# Third party QC of individual projects

# 1. Analysis of overall trend

# 1. Analysis of overall trend

**Average, dispersion and standard deviation for 50 targeted projects (standardized score) (and the results in 2018)**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.903 | 0.957 | 0.901 | 0.894 | 0.913 | 0.968 | 0.905 | 0.955 |
| Dispersion | 0.005 | 0.003 | 0.003 | 0.006 | 0.004 | 0.000 | 0.005 | 0.003 |
| Standard deviation | 0.072 | 0.051 | 0.051 | 0.077 | 0.067 | 0.022 | 0.068 | 0.051 |

Note: In 2018, DPs were included in the TCs in data aggregation. In 2019, DPs were separately calculated, so the data for 2018 was also calculated again.
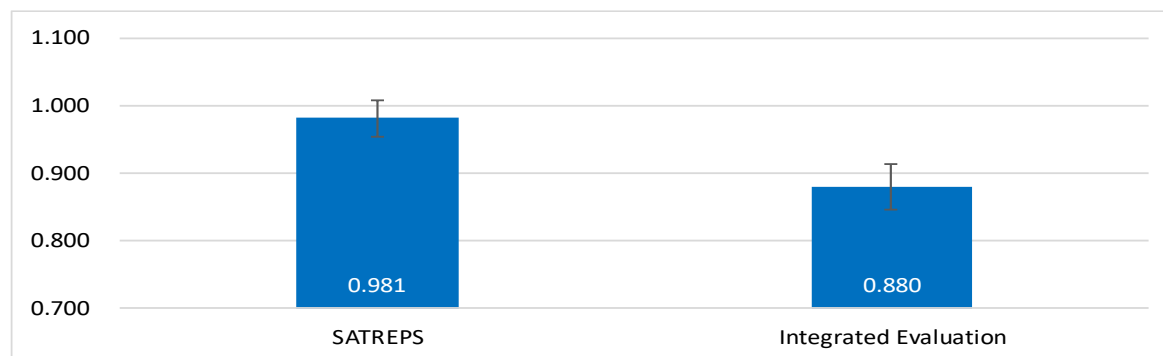


Note：Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

● The average DP score is low. The reason is that the number of DP projects is small and their standardized scores are relatively low.

# 1. Analysis of overall trend

**Average, dispersion and standard deviation for 2 SATREPS and 2 integrated evaluation**

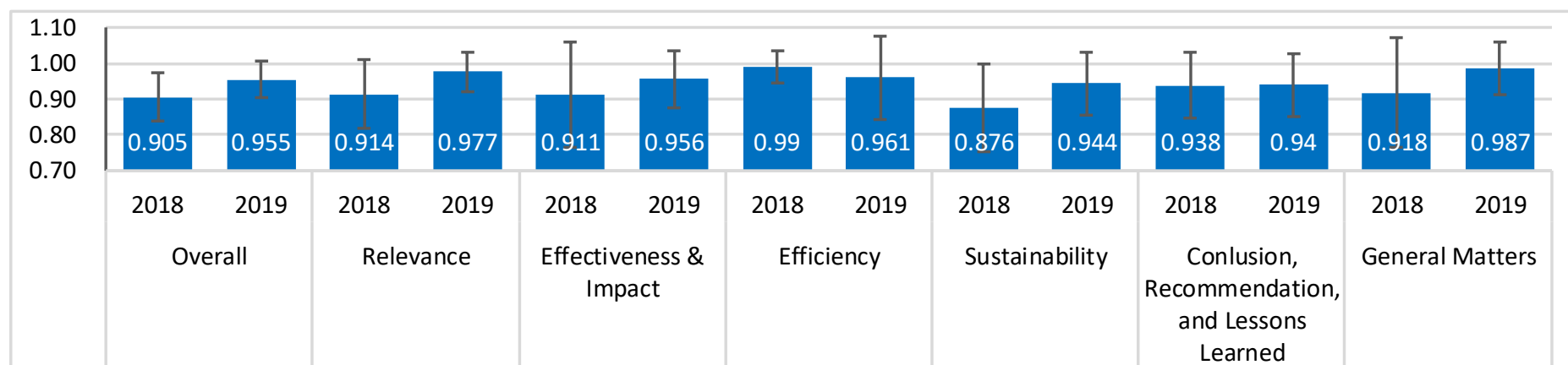| | SATREPS | Integrated evaluation |
|---|---|---|
| Average | 0.981 | 0.880 |
| Dispersion | 0.001 | 0.001 |
| Standard deviation | 0.027 | 0.033 |



Note ：Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

- The targeted projects in 2019 included two SATREPS projects and two integrated evaluation (one evaluation of a two-phase technical cooperation projects). The average, dispersion, and standard deviation of these projects are as described above.

- For SATREPS, the third party QC result was almost the perfect score (0.981), and the average score of about 90% (0.880) was secured for the Integrated evaluation. It can be said that the quality of both types was assured as in the other targeted projects.

# 1. Analysis of overall trend

**Average, dispersion and standard deviation for 50 targeted projects (standardized score) (and the results in 2018)**

| | Overall | | Relevance | | Effectiveness / Impact | | Efficiency | | Sustainability | | Conclusion, Recommendation, and Lessons Learned | | General Matters | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.905 | 0.955 | 0.914 | 0.977 | 0.911 | 0.956 | 0.990 | 0.961 | 0.876 | 0.944 | 0.938 | 0.940 | 0.918 | 0.987 |
| Dispersion | 0.005 | 0.003 | 0.009 | 0.003 | 0.022 | 0.006 | 0.002 | 0.014 | 0.015 | 0.008 | 0.008 | 0.008 | 0.024 | 0.005 |
| Standard deviation | 0.068 | 0.051 | 0.097 | 0.056 | 0.148 | 0.081 | 0.044 | 0.117 | 0.124 | 0.089 | 0.092 | 0.088 | 0.155 | 0.074 |



| 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.905 | 0.955 | 0.914 | 0.977 | 0.911 | 0.956 | 0.99 | 0.961 | 0.876 | 0.944 | 0.938 | 0.94 | 0.918 | 0.987 |
| Overall | | Relevance | | Effectiveness & Impact | | Efficiency | | Sustainability | | Conlusion, Recommendation, and Lessons Learned | | General Matters | |

Note：Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

# 1. Analysis of overall trend

**Correlation coefficient of the standardized score of each evaluation criteria to the entire standardized score**

The correlation coefficient of the standardized score of each evaluation criteria with respect to the standardized score of the entire standardized score was calculated for all 50 subjects, and the following results were obtained.

| | Correlation coefficient of the standardized score of each evaluation criteria to the entire standardized score | Number of third party QC items |
|---|---|---|
| Overall | 1.000 | 35 |
| Relevance | 0.597 | 6 |
| Effectiveness / Impact | 0.514 | 7 |
| Efficiency | 0.256 | 4 |
| Sustainability | 0.845 | 11 |
| Conclusion, Recommendation, and Lessons Learned | 0.370 | 4 |
| General Matters | 0.216 | 3 |

[Level of correlation coefficient]
0.7-1 fairly strong correlation
0.4-0.7 strong correlation
0.2-0.4% weak correlation
0-0.2 almost no correlation

[Calculation formula of correlation coefficient (reference)]
Correlation coefficient between x and y = (Covariance of x and y) / (standard deviation of x) x (standard deviation of y)

- A positive correlation was observed in all criteria. A slightly strong correlation was found for "Relevance" and "Effectiveness / Impact", and a fairly strong correlation was found for "Sustainability".

- Basically, the more the number of third party QC items, the stronger the correlation. This can be said to be a natural result in totaling.

- A particularly strong correlation was observed in "Sustainability". Basically, in third party QC, it is checked that if the sustainability of each aspect (political, organizational, technical, and financial) is evaluated based on the data and information necessary for analysis. It is assumed that evaluators who analyze sustainability properly have high overall accuracy of analysis, and conversely, evaluators who have insufficient sustainability analysis have low overall accuracy of analysis.

# 1. Analysis of overall trend

- The average of the standardized scores for the 50 targeted projects is 0.955. It increased from 0.905 points in 2018. As in the previous year, the third party QC scoring results were high, and almost all items were judged as "yes".

- The rise in the standardized score was due to the fact that many internal evaluation were assured of quality, same as in 2018. In addition, this year, the scoring standards were reviewed and the standards were clarified.

- The third party QC scoring standard emphasized the confirmation of formal requirements so as not to be influenced by arbitrary judgment. For items that are difficult to judge, it was tried to minimize subjectively lowering the score as much as possible from the viewpoint of uniform scoring standards. As a result, it became easier to get a slightly higher score.

- By scheme, GA and TC averaged almost the same level. The average score of the DP was slightly lower. The number of DP projects was small, and overall, score was slightly lower than other schemes.

- By the evaluation criteria (OECD-DAC 5 evaluation criteria), the average score of "Efficiency" was higher in 2018 and the others were slightly lower. On the other hand, in 2019, it was secured around 0.950 points for all criteria. The factors are the same as described in the factors for the rise of the standardized score. The average scores for "Sustainability" and "Conclusions, Recommendation, and Lessons Learned" were slightly lower than others.

- In 2018, "efficiency" had a high average score and small variations. In addition, "Effectiveness / Impact" and "Sustainability" had low average scores and large variations. In 2019, these kind of trends were alleviated, and there was no remarkable trend in the scoring results among each evaluation criteria.

# 1. Analysis of overall trend

**Evaluation result**

| | Highly satisfactory | Satisfactory | Partially satisfactory | Unsatisfactory |
|---|---|---|---|---|
| Number of applicable projects | 17 | 20 | 9 | 4 |
| Standardized score (average) | 0.973 | 0.945 | 0.932 | 0.973 |
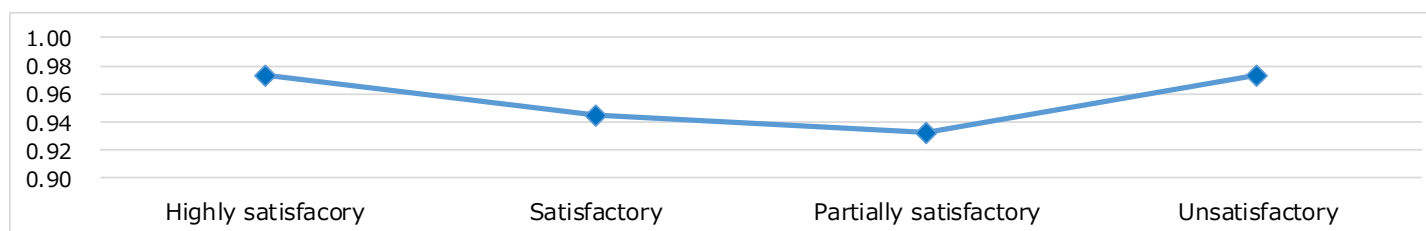


- As for the classification of evaluation results, the standardized score of "highly satisfactory" and "unsatisfactory" were slightly higher, and the standardized scores of "satisfactory" and "partially satisfactory" were slightly lower. When a clear evaluation result ("highly satisfactory" or "unsatisfactory") is derived, the basis for judgment and information tend to be clear, while intermediate evaluation results ("satisfactory" or "partially satisfactory") may be somewhat ambiguous and may result in differences in the scoring of third party QC.

# 1. Analysis of overall trend

**The difference between the results of self-assessment and third party QC was analyzed. To be specific, for 50 projects where self-assessment was conducted, the difference between the standardized scores of self-assessment and third party QC was analyzed. In each of the items of "Relevance", "Effectiveness / Impact", "Efficiency", "Sustainability", "Conclusions, Recommendations, and Lessons Learned", and "General Matters", the extent to which differences occurred was analyzed.**

**Comparison of third party QC results and self-assessment results of 50 targeted projects by standardized scores**

|  | Overall | Relevance | Effectiveness / Impact | Efficiency | Sustainability | Conclusion, Recommendation, and Lessons Learned | General Matters |
|---|---|---|---|---|---|---|---|
| Third party QC | 0.955 | 0.977 | 0.956 | 0.961 | 0.944 | 0.940 | 0.987 |
| Self-assessment | 0.938 | 0.980 | 0.966 | 0.880 | 0.912 | 0.968 | 0.897 |

- In the overall, the standardized score of third party QC exceeded the standardized score of self-assessment. It can be said that the third party QC considered the quality of the evaluation higher. However, the third party QC, unlike self-assessment, is difficult to make subjective judgments and mainly focuses on confirmation of formalities, and as a result, it seems that the score tends to be slightly higher.

- As for "Efficiency", the result of self-assessment was lower. In the third party QC, there were many checks of "not applicable", and when applicable, there were many checks of "yes". In the self-assessment, each item was scored a bit stricter.

- As for "General Matters", the result of self-assessment was lower. For example, for items related to the cases "when there are important restrictions and points to be considered", slightly stricter scoring was performed in the self-assessment. In third party QC, items that are difficult to judge from the appearance of the evaluation result sheet have been checked in "yes" more frequently. In self-assessment, the score is likely to be more rigorous than third party QC because the self-assessment will judge whether the item is sufficient or insufficient based on its own experience.

# 1. Analysis of overall trend

**Gap analysis of checklist items between self-assessment and third party QC**

| Items | Number of items |
|---|---|
| Third party QC = Self-assessment | 902 (54.9%) |
| Third party QC > Self-assessment | 81 (4.9%) |
| Third party QC < Self-assessment | 76 (4.6%) |
| The number of items where "not applicable" was checked by self-assessment and/or third party QC, the number of items in the third party QC that are not included in the self-assessment items, and the number of items not answered by self-assessment | 583 (35.5%) |

the population of gap analysis is calculated as:  34 TC x 33 items + 4 DP x 31 items + 12 GA x 33 items = 1,642.

- The number of items in case that third party QC = self-assessment: more than 50%,

- The number of items in case that "not applicable", items with no questions in self-assessment or items not answered in self-assessment: less than 40%

- The number of items with gaps was about 5% each for third party QC > self-assessment and third party QC < self-assessment.

# 1. Analysis of overall trend

| Evaluation criteria | Gap between self-assessment and third party QC | Number of items (rate to the population) | Comment |
|---|---|---|---|
| Relevance (34TC×6items＋4DP×6items＋12GA×6items=300items) | Third party QC = self-assessment | 186 (62.0%) | ● The results of self-assessment and third party QC are generally consistent. |
| | Third party QC > self-assessment | 7 (2.3%) | |
| | Third party QC < self-assessment | 7 (3.0%) | |
| | NA, with no items, with no answer | 98 (32.7%) | |
| Effectiveness / Impact (34TC×7items＋4DP×5items＋12GA×6items=330items) | Third party QC = self-assessment | 221 (67.0%) | ● The number of cases such as NA is small.<br>● The difference between the third party QC and the self-assessment was slightly more than other items (about 10%).<br>● This is probably due to the fact that there are many items determined by the third party quality checker / self-assessor, not just a format check such as the presence or absence of information and the appropriateness of the judgment. |
| | Third party QC > self-assessment | **13 (3.9%)** | |
| | Third party QC < self-assessment | **20 (6.1%)** | |
| | NA, with no items, with no answer | 76 (23.0%) | |
| Efficiency (34TC×3items＋4DP×3items＋12GA×4items=162items) | Third party QC = self-assessment | 54 (33.3%) | ● The number of cases such as NA was very large.<br>● In many evaluations, only evaluation judgment is made based on comparison between plan and actual, and many of the items in the case of "a significant difference between the plan and the actual" and "evaluation judgment according to increase / decrease in output" were checked as NA. |
| | Third party QC > self-assessment | 8 (4.9%) | |
| | Third party QC < self-assessment | 6 (3.7%) | |
| | NA, with no items, with no answer | **94 (58.0%)** | |

# 1. Analysis of overall trend

| Evaluation criteria | Gap between self-assessment and third party QC | Number (rate to the population) | Comment |
|---|---|---|---|
| Sustainability (34TC×10items＋4DP×10items＋12GA×10items=500items) | Third party QC = self-assessment | 251 (50.2%) | ● The difference between the third party QC and self-assessment was slightly more than other items (about 13%). ● This is probably due to the fact that there are many items determined by the third party quality checker / self-assessor, not just a format check such as the presence or absence of information and the appropriateness of the judgment. |
| | Third party QC > self-assessment | **37 (7.4%)** | |
| | Third party QC < self-assessment | **27 (5.4%)** | |
| | NA, with no items, with no answer | 185 (37.0%) | |
| Conclusion, Recommendation, and Lessons Learned (34TC×4items＋4DP×4items＋12GA×4items=200items) | Third party QC = self-assessment | 119 (59.5%) | ● The ratio of third party QC < self-assessment was slightly higher. ● In the self-assessment, it seems that the relevance between the recommendations and lessons learned and the evaluation results was somewhat strictly checked. |
| | Third party QC > self-assessment | 5 (2.5%) | |
| | Third party QC < self-assessment | **14 (7.0%)** | |
| | NA, with no items, with no answer | 62 (31.0%) | |
| General Matters (34TC×3items＋4DP×3items＋12GA×3items=150items) | Third party QC = self-assessment | **71 (47.3%)** | ● The ratio of third party QC < self-assessment was slightly higher. ● There were many cases such as NA. Most of the items in the case of "sample survey" and "important restrictions and considerations" were checked as NA. |
| | Third party QC > self-assessment | **11 (7.3%)** | |
| | Third party QC < self-assessment | 0 (0.0%) | |
| | NA, with no items, with no answer | 68 (45.3%) | |

# 2. Analysis by evaluation criteria

# 2. Analysis by evaluation criteria (Relevance)

**Average, dispersion, and standard deviation of "Relevance" based on standardized scores of 50 targeted project**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.922 | 0.970 | 0.929 | 0.975 | 0.891 | 1.000 | 0.914 | 0.977 |
| Dispersion | 0.009 | 0.004 | 0.008 | 0.003 | 0.010 | 0.000 | 0.009 | 0.003 |
| Standard deviation | 0.096 | 0.065 | 0.091 | 0.050 | 0.099 | 0.000 | 0.097 | 0.056 |

✓ There is no particular tendency for each scheme.

✓ Most of the descriptions are without problems in general.
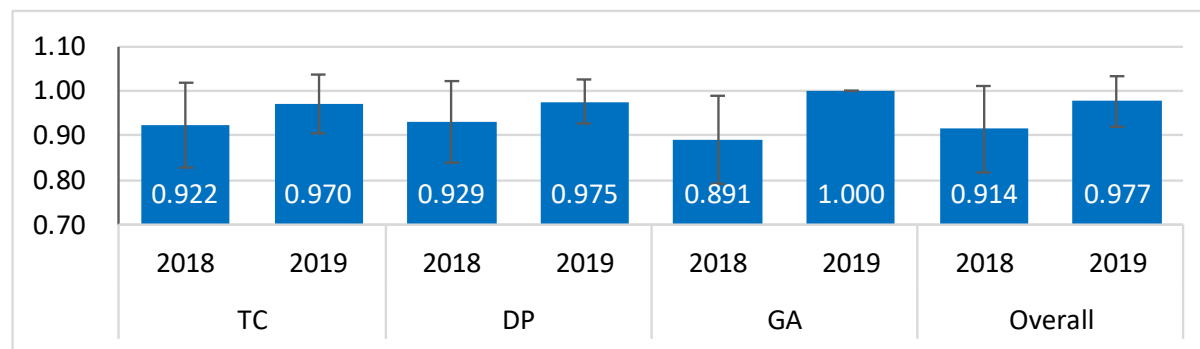


Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

✓ Regarding the consistency with Japan's ODA policy, in some evaluations, explanations on relevant objectives were insufficient, and it was difficult to understand how Japan's ODA policy was consistent with projects.

✓ Regarding the adequacy of the approach, there were few evaluations in which "effectiveness / impact" and "sustainability" were judged to be "low", and ticks for "not applicable" accounted for the majority.

# 2. Analysis by evaluation criteria (Relevance)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 1 | [Development policy] Does the evaluation report mention (1) the names of the development policies of the country, (2) the years covered by these policies, and (3) clarification of the project's position in these policies at the time of ex-ante evaluation, and (4) does the evaluation judgement align with them? | | | TC: 1.971 DP: 2.000 GA: 2.000 | [Example of Comment] ✓ Insufficient explanation of the position of the project in the development policy. ✓ Because it is the integrated evaluation, it is necessary to conduct an analysis to confirm the consistency with the partner country's development policy at the time of completion of Phase 1 and planning of Phase 2. [Analysis] ✓ Most of the descriptions are with no problem in general. |
| 2 | [Development policy] Does the evaluation report mention (1) the names of the development policies of the country, (2) the years covered by these policies, and (3) clarification of the project's position in these policies at the time of project completion, and (4) does the evaluation judgement align with them? | [Development policy] Does the evaluation report mention (1) the names of the development policies of the country, (2) the years covered by these policies, and (3) clarification of the project's position in these policies at the time of ex-post evaluation, and (4) does the evaluation judgement align with them? | | TC: 1.971 DP: 2.000 GA: 2.000 | [Example of Comment] ✓ Insufficient explanation of the position of the project in the development policy. ✓ Because it is the integrated evaluation, it is necessary to conduct an analysis to confirm the consistency with the partner country's development policy at the time of completion of Phase 1 and planning of Phase 2. [Analysis] ✓ Most of the descriptions are with no problem in general. |
| 3 | [Development needs] Does the evaluation report mention (1) the development needs of the country, (2) the information to indicate the existence of the needs (e.g., description of the context of the project, etc.) at the times of both ex-ante evaluation and project completion, and (3) does the evaluation judgement align with them? | [Development needs] Does the evaluation report mention (1) the development needs of the country, (2) the information to indicate the existence of the needs (e.g., description of the context of the project, etc.) at the times of both ex-ante evaluation and ex-post evaluation, and (3) does the evaluation judgement align with them? | | TC: 1.912 DP: 2.000 GA: 2.000 | [Example of Comment] ✓ It is difficult to understand what kind of development needs was judged based on what kind of basis information. ✓ Tense is unclear. ✓ There is no explanation of development needs at the time of project completion. [Analysis] ✓ Most of the descriptions are with no problem in general. (Many evaluation sheets explains development needs in detail in the section of "background".in evaluation result sheet) |

# 2. Analysis by evaluation criteria (Relevance)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 4 | [Japan's ODA policy] Does the evaluation report mention (1) the names of the ODA policies for the country, (2) the years covered by these policies, and (3) relevant objectives, etc., at the time of ex-ante evaluation, and (4) does the evaluation judgement align with them? | | | TC: 1.882 DP: 1.750 GA: 2.000 | [Example of Comment]<br>✓ Insufficient explanation on how Japan's ODA policy is consistent with the project.<br>✓ Insufficient explanation of related objectives.<br>✓ There is no description of the targeted year.<br>✓ In principle, it should be consistent with the same policy referred at the time of the ex-ante evaluation (such as the country's assistance policy by MOFA).<br>[Analysis]<br>✓ Explanations on related objectives were sometime insufficient, and it was difficult to understand how Japan's ODA policy was consistent with the project.<br>✓ In addition to simply indicating the priority areas of the country-specific development cooperation policy, it should be stated why the project is judged to be consistent with the priority areas. |
| 5 | [Appropriateness] In case the project experienced a problem for Effectiveness/Impact and/or Sustainability, does the evaluation report reasonably analyze the appropriateness of the project based on the evidence? | | | TC: 1.750 DP: 2.000 GA: All are ticked in "Not Applicable" | [Example of Comment]<br>✓ Descriptions are inconsistent with lessons learned (The lessons learned states that the approach of the project was not appropriate, but the relevance judged that the approach was appropriate)<br>[Analysis]<br>✓ There were few evaluations that judged Effectiveness / Impact and Sustainability to be "low", and "not applicable" accounted for the majority. |
| 6 | [Evaluation Result] Is the evaluation judgment on Relevance as a whole consistent with the judgments on (1) the development policy, (2) the development needs and (3) Japan's ODA policy (and, if applicable, (4) appropriateness of the approach)? | | | TC: 1.971 DP: 2.000 GA: 2.000 | [Example of Comment]<br>✓ The problem in the appropriateness of the approach is indicated, but is not considered in the evaluation judgment.<br>[Analysis]<br>✓ Most of the descriptions are with no problem in general. |

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

**Average, dispersion, and standard deviation of "Effectiveness and Impact" based on standardized scores of 50 targeted project**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.894 | 0.948 | 0.988 | 1.000 | 0.917 | 0.963 | 0.911 | 0.956 |
| Dispersion | 0.028 | 0.007 | 0.001 | 0.000 | 0.013 | 0.006 | 0.022 | 0.006 |
| Standard deviation | 0.169 | 0.086 | 0.029 | 0.000 | 0.114 | 0.077 | 0.148 | 0.081 |

✓ The high average score of the DP seems to be due to the fact that it is easy to clearly determine the achievement of output and outcomes in DP that often produce some kind of master plan.
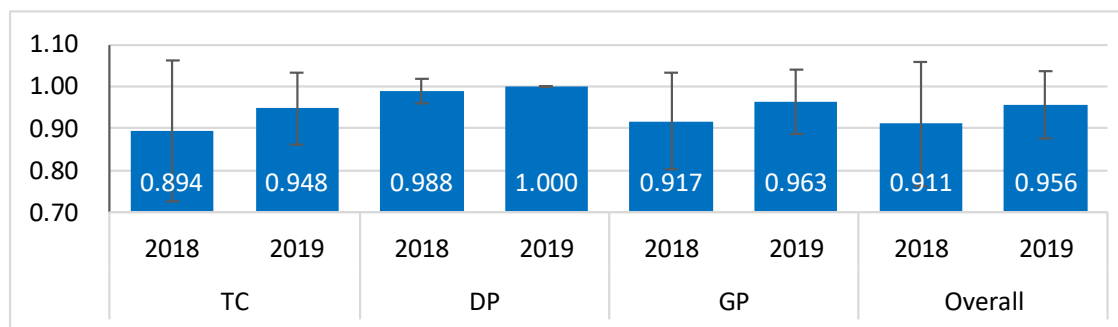


Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

✓ Some projects have problems in setting indicators (no target set, ambiguous, cannot measure the achievement of target, etc.)

✓ Regarding the Overall Goal / Impact, the TC scored slightly lower. At the time of PDM creation, there are many cases where indicators were originally set for which it is difficult to obtain data, which seems to have affected the difficulty of evaluation.

✓ Regarding the treatment of inappropriate indicator setting, it was difficult to judge how much the supplementary information should be examined and analyzed. As a result, scoring was mainly conducted for cases actually using the supplementary information, and "not applicable" increased for other cases.

✓ It is difficult to judge the necessity of analysis for external factors. Many evaluation results do not state any clear external factors. In that case, it is difficult to judge whether or not an external factor actually exists. As a result, the score of "not applicable" increased.

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 7 | [Project Purpose] (1) Does the evaluation report present and properly analyze the indicator data through the time of project completion, and (2) are they consistent with the judgement of "achieved, partially achieved, or unachieved"? | [Achievement of the Objectives at the timing of project completion] Does the evaluation report present concrete evidence and properly analyze the status of delivered output and the achievement of the project purpose at the time of project completion? | [Effectiveness] With respect to the project objective(s), (1) does the evaluation report present and analyze the indicator data for the period between the project completion and the ex-post evaluation (including the target year)? (2) If the report fails to include certain indicator data, does it offer an analysis based on supplemental information designated in the EFM? (3) Does the report present the information on qualitative effects (including effects of the soft component if applicable)? (4) Does the evaluation judgement of Effectiveness consistent with them? | TC: 1.939 DP: 2.000 GA: 1.833 | [Example of Comment]<br>✓ There is no data at the time of ex-post evaluation.<br>✓ The reason for determining that the indicator was achieved is unclear.<br>✓ (Issues in setting indicators) Target are not set, and definitions of indicators are ambiguous. As a result, the persuasion of evaluation judgment is weak.<br>✓ (Issues in setting indicators) For seeing the effect of the project, the XX rate should be analyzed instead of the XX number.<br>✓ Since only data after the completion of the project is described, it is not possible to judge whether the indicators would be improved by the project.<br>✓ The number of years of data is small, and it is difficult to measure the degree of improvement.<br>✓ It is doubtful that the indicated data alone shows the project effect. It seems to be necessary to analyze supplementary information.<br>✓ Data credibility is slightly lacking.<br>✓ [GA] Soft Component is being implemented, but there is no description of its effects.<br><br>[Analysis]<br>✓ The comment example is as described above.<br>✓ In addition, there are some projects having a problem with the setting indicators (no target set, ambiguous, target achievement could not be measured, etc.). In such a case, it is desirable to consider using the supplementary information, but it is difficult to make a judgement by third party QC how much analysis or examination of supplementary information is required . |

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 8 | [Continuation Status of Project Effects] (1) Does the evaluation report present and properly analyze the data showing continued effects through the time of ex-post evaluation? (2) Does the judgement of "continued, partially continued or not continued" consistent with them? (3) If the report fails to include certain indicator data, does it offer an analysis based on supplemental information designated in the EFM? | No Checklist item | [Effectiveness] With respect to the operation status of the facilities/equipment developed under the project and generation/continuation status of projects effects through the time of ex-post evaluation, (1) does the evaluation report present and analyze quantitative data such as capacity utilization, etc. or qualitative information? (2) Do they accordantly reflect on the evaluation judgement? | TC: 1.882<br>DP: -<br>GA: 1.917 | [Example of Comment]<br>✓ There is no data at the time of ex-post evaluation.<br>✓ The judgement in the continuation of the effect is doubtful.<br>✓ Since the data at the time of the terminal evaluation is used, the continuation status at the time of the ex-post evaluation is not verified.<br>✓ Data is little and analysis is insufficient. |

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 9 | [Overall Goal] (1) Does the evaluation report present and correctly analyze the indicator data from the time of project completion to the time of ex-post evaluation including the target year? (2) Are the consistent with evaluation judgement? (3) If the report fails to include certain indicator data, does it offer an analysis based on supplemental information designated in the EFM? | [Utilization level of the proposed plan] (1) Does the evaluation report present and correctly analyze the indicator data showing the status of Utilization level about the proposed plan r? (2) If the report fails to include certain indicator data, does it offer an analysis based on supplemental information designated in the EFM? | [Impact] With respect to the expected impact (the part after "thereby contributing to" in "Objectives of the Project"), (1) does the evaluation report present and analyze quantitative data or qualitative information through the time of ex-post evaluation? (2) Do the accordantly reflect on the evaluation judgement? | TC: 1.676<br>DP: 2.000<br>GA: 1.917 | [Example of Comment]<br>✓ There is no data at the time of ex-post evaluation.<br>✓ The judgement in the continuation of the effect is doubtful.<br>✓ Achievement of indicators cannot be judged only with the data described.<br>✓ (Issues in setting indicators) The definition of indicators is ambiguous, and it is difficult to understand what can determine the achievement of indicators.<br>✓ There are not the specific data and time series data indicated.<br>✓ There are many qualitative indicators, but sufficient evidence is not provided.<br>✓ There is insufficient information to derive evaluation judgments using only the comments of related parties.<br>✓ [GA] There is no statement that the expected impact has been achieved, and it makes difficult to understand the basis of the evaluation judgement.<br>[Analysis]<br>✓ The comment example is as described above.<br>✓ In addition, there are some projects having a problem with the setting indicators (no target set, ambiguous, target achievement could not be measured, etc.). In such cases, it is desirable to consider using the supplementary information, but it is difficult to make a judgement by third party QC how much analysis or examination of supplementary information is required<br>✓ The score of TC is slightly low. At the time of PDM creation, there are many cases where indicators for which data are originally difficult to obtain are set, and it makes evaluation difficult. |

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 10 | [Project Purpose & Overall Goal] In cases as there are problems in indicators to be verified (such as no target value, inappropriate indicator for continuation status of project effects, vague definition of indicator ), is the project judged "Not verified", judged based on supplemental information or judged with reasonable explanations? | No Item | | TC: 1.850<br>DP: -<br>GA: - | [Example of Comment]<br>✓ The judgment in some indicators is doubtful. (it is judged as "unverifiable", but it should be judged as "XXX", etc.)<br>[Analysis]<br>✓ As described in the analysis of project purpose and overall goal, it was difficult to judge how much the supplementary information should be examined and analyzed. As a result, scoring was mainly conducted for cases actually using the supplementary information, and "not applicable" increased. |
| 11 | Are the grounds for judging individual sub-items of Evaluation/Impact objective and neutral (i.e., Is the judgment not made only based on an opinion of one person or impression? Is the objective not judged as achieved without reasonable explanations even though the indicator values substantially fall short of target values? Is the objective not judged as achieved even though the target values or actual values are unavailable? No it isn't => tick "Yes"; Yes it is => tick "No") | | | TC: 1.971<br>DP: 2.000<br>GA: 2.000 | [Example of Comment]<br>✓ Judgment is somewhat unclear. |
| 12 | In case that the factors other than the project mainly affected achievement/non-achievement on individual sub-items and for the increase/decrease in indicator values, does the evaluation report state reasons for it? | | | TC: 2.000<br>DP: All are ticked in "Not Applicable"<br>GA: 2.000 | [Analysis]<br>✓ It is difficult to judge whether external factors need to be analyzed. As a result, all the ticks were in "yes" or "not applicable" because scoring were only made for cases where external factors were actually described. |

# 2. Analysis by evaluation criteria (Effectiveness / Impact)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 13 | [Evaluation Result] Is the evaluation judgment on Effectiveness/Impact as a whole consistent with the judgments on 1. the degree to which Project Purpose has been achieved, 2. the status of the continued effects of Project Purpose toward achieving Overall Goal, and, 3. the degree to which Overall Goal has been achieved? | [Evaluation Result] Is the evaluation judgement on Effectiveness/Impact as a whole consistent with the judgements on 1. the degree of the achievement of the objectives by the project completion, and 2. the status of utilization of the proposed plan by the time of ex-post evaluation. | [Evaluation Result] Is the evaluation judgment on Effectiveness/Impact as a whole consistent with the judgments on 1. the quantitative and qualitative effects (Effectiveness) and 2. the expected impact. 3. If other positive/negative impacts are stated and they are considered significant, does the evaluation judgment reflect them? | TC: 2.000<br>DP: 2.000<br>GA: 1.909 | [Example of Comment]<br>✓ Question on evaluation judgment (evaluation judgment is too severe)<br>✓ Description of Impact is not originally expected impact but other positive or negative impact. If other positive or negative impacts are included in the evaluation of effectiveness / impact, it is necessary to explain the detail of the analysis.<br>[Analysis]<br>✓ Most of the descriptions are with no problem in general. |

# 2. Analysis by evaluation criteria (Efficiency)

**Average, dispersion, and standard deviation of "Efficiency" based on standardized scores of 50 targeted project**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.993 | 0.978 | 1.000 | 0.792 | 0.979 | 0.969 | 0.990 | 0.961 |
| Dispersion | 0.002 | 0.009 | 0.000 | 0.063 | 0.003 | 0.006 | 0.002 | 0.014 |
| Standard deviation | 0.041 | 0.095 | 0.000 | 0.250 | 0.055 | 0.078 | 0.044 | 0.117 |

✓ The average score of the DP is low and the standard deviation (variation) is large. In the DP, the number of the targeted projects is small and low scores greatly affected the overall average score.
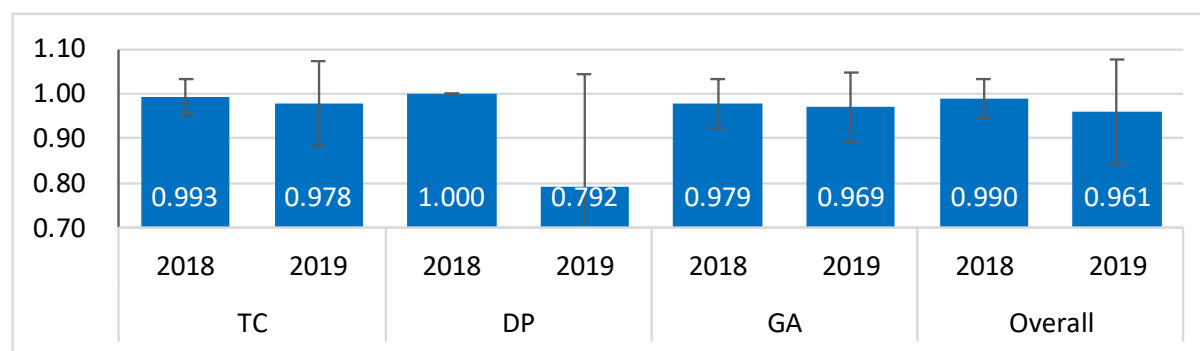


Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

✓ When performing simple comparisons of plan and actual results, most of the evaluations are with no problem. On the other hand, when there is a large discrepancy between the plan and the actual, or when the evaluation is based on the increase or decrease of the output instead of the simple comparison between the plan and the actual, there are more number of evaluations with insufficient explanation of the reason of judgement.

✓ Regarding the projects having large discrepancy between the plan and actual, there were more number of evaluations where the reason is described but is not reasonable.

✓ In cases where evaluation judgments were made based on changes in output, scoring was deducted in cases where the reason used to derive the judgments is not sufficiently explained.

# 2. Analysis by evaluation criteria (Efficiency)

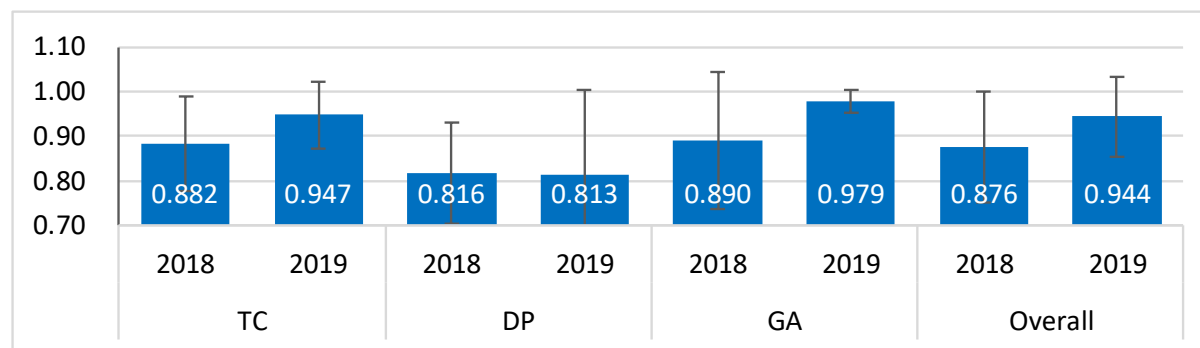| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 14 | No Item | | In case there is a significant gap between the planned and actual outputs, does the evaluation report mention it along with reason(s) in "Efficiency" or other sections (such as "Project Outline" and "Effectiveness/Impact")?<br><br>(1. Check 1. whether it is described. 2. its rationality) | TC: -<br>DP: -<br>GA: 2.000 | [Analysis]<br>The checklist items are only for GA, and there were many checks for "not applicable". |
| 15 | In case there was a significant discrepancy between the planned and actual inputs (project period and project cost), (1) does the evaluation report mention the reason(s), and (2) is it rational? | | | TC: 1.800<br>DP: 2.000<br>GA: 1.833 | [Example of Comment]<br>✓ Although the reason for the increase in output is described, it is necessary to explain its background in more detail.<br>✓ Although there is an explanation of the change from the planned input, there is no description on the reason or verification of the rationality of the reason.<br>[Analysis]<br>✓ "Not applicable" are ticked in many projects.<br>✓ In some cases, the reasons are stated but no explanation is given as to their rationality. |

# 2. Analysis by evaluation criteria (Efficiency)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 16 | If the assessment of the inputs is made based on the increase or decrease in the outputs rather than on a simple comparison between the planned and actual inputs, (1) does the evaluation report explain it in relation to the outcomes? (2) Is the explanation reasonable? | | | TC: 1.500<br>DP: 1.500<br>GA: 2.000 | [Example of Comment]<br>✓ If it is judged that the efficiency is high, it is necessary to explain in more detail how much output has been added.<br>✓ There is no mention of the relationship with the outcome.<br>[Analysis]<br>✓ "Not applicable" are ticked for many projects.<br>✓ Scoring by third party QC is deducted in cases where the explanation of the concept and rationality used to derive the evaluation judgment was insufficient. |
| 17 | [Evaluation Result] Is the evaluation judgment on Efficiency as a whole consistent with the judgments on (1) the project period and (2) the project cost? | | | TC: 2.000<br>DP: 1.250<br>GA: 1.917 | [Example of Comment]<br>✓ Questioning in the evaluation judgment (Is "XXX" instead of "XXX")<br>✓ When determining "XXX" in consideration of an increase in output, detailed explanation of such judgement is necessary.<br>[Analysis]<br>✓ In general, most of the evaluations are without major problem.<br>✓ The average score of the DP is low because the number of the entire DP projects is small and the evaluation of the low score had a large effect on average score. |

# 2. Analysis by evaluation criteria (Sustainability)

**Average, dispersion, and standard deviation of "Sustainability" based on standardized scores of 50 targeted project**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.882 | 0.947 | 0.816 | 0.813 | 0.890 | 0.979 | 0.876 | 0.944 |
| Dispersion | 0.011 | 0.006 | 0.013 | 0.036 | 0.024 | 0.001 | 0.015 | 0.008 |
| Standard deviation | 0.105 | 0.075 | 0.114 | 0.189 | 0.155 | 0.026 | 0.124 | 0.089 |

✓ The average score of DP is low and the standard deviation (variation) is large. In the DP, the number of projects is small and low scores greatly affected the average score.



Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

✓ Regarding GA, many evaluations are well described. Regarding TC and DP, various factors are involved in the analysis of sustainability, so it seems more difficult to derive evaluation judgement. On the other hand, in GA, basically, the viewpoint of whether the constructed facilities and the procured materials and equipment were used appropriately is a major part of judgments. It seems that the analysis is somewhat simpler than other two schemes.

✓ As for technical sustainability, the average score was lower in TC and DP. Analyzing the sustainability of these schemes from a technical point of view seems to be rather complicated and it is not easy to derive reasonable evaluation judgments.

✓ Regarding factors and prospects, the average score was generally low. In many evaluations, the factors of the issues are described, but not the prospects of the issues in the future.

# 2. Analysis by evaluation criteria (Sustainability)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 18 | [Policy Aspect] Does the evaluation report mention (1) the names of the development policies of the country, (2) the target years of the policies, and (3) relevant objectives, etc., at the time of ex-post evaluation? (4) Are they consistent with evaluation judgement? | No Item | | TC: 1.971 DP: 1.500 GA: - | [Example of Comment] ✓ Relevant objectives are not described. ✓ It is a policy for the period of XXXX (year), not a policy at the time of ex-post evaluation. ✓ The name of the policy is not specified. ✓ The policy is a bit old. It was necessary to explain whether the policy was effective at the time of the ex-post evaluation. [Analysis] ✓ In general, most of the evaluations are without major problem. |
| 19 | [Institutional Aspect] Does the evaluation report mention (1) the organization(s) responsible for sustaining the project effects, and (2) make the roles of such organization(s) in sustaining the project effects clear? | [Institutional Aspect] Does the evaluation report mention (1) the organization(s) responsible for promoting the proposed development plan, and (2) make the roles of such organization(s) in promoting the proposed development plan clear? | [Institutional Aspect] Does the evaluation report mention (1) the organization(s) responsible for sustaining the project effects, and (2) make the roles of such organization(s) in sustaining the project effects clear? | TC: 1.941 DP: 1.750 GA: 2.000 | [Example of Comment] ✓ Insufficient explanation of the role of the organization responsible for sustaining the project effects. [Analysis] ✓ In general, most of the evaluations are without major problems. ✓ The average score of the TC and DP are slightly lower than GA. |

# 2. Analysis by evaluation criteria (Sustainability)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 20 | [Institutional Aspect] Does the evaluation report (1) make it clear the reported organizational structure is necessary for what specific activities/services (e.g., dissemination of the model) ? (2) Does the evaluation report present the data about the number of staff, and (3) the opinion of concerned party? | [Institutional Aspect] Does the evaluation report (1) make it clear the reported organizational structure is necessary for what specific activities to promote the proposed development plan? (2) Does the evaluation report present the data about the number of staff, and (3) the opinion of concerned party? | [Institutional Aspect] Does the evaluation report present (1) the data of the organizational structure and the number of staff, and (2) opinion of concerned party? | TC: 1.971<br>DP: 2.000<br>GA: 2.000 | [Example of Comment]<br>✓ Data such as the number of staffs is not described.<br>[Analysis]<br>✓ In general, most of the evaluations are without major problem. |
| 21 | [Institutional Aspect] With respect to the grounds on which the institutional levels such as the organizational structure and the number of staff are judged to be sufficient/insufficient, (1) does the evaluation report show data in the explanation? (2) Is the judgement consistent with analysis? | | | TC: 1.971<br>DP: 2.000<br>GA: 2.000 | [Example of Comment]<br>✓ Insufficient analysis for evaluation judgment.<br>[Analysis]<br>✓ In general, most of the evaluation are without major problem. |
| 22 | [Technical Aspect] Does the evaluation report (1) make it clear the reported technique is necessary for what specific activity/service (e.g., dissemination of the model)? (2) Does the evaluation report present the data for the grounds on which it judges the technical level to be sufficient/insufficient,  and (3) opinion of concerned party?? | [Technical Aspect] Does the evaluation report show (1) the data of technique, and (2) make it clear the reported technique is necessary for what specific activity to promote the proposed development plan? Does the evaluation report present the data for the grounds on which it judges the technical level to be sufficient/insufficient, and opinion of concerned party? | [Technical Aspect] Does the evaluation report present (1) the data for the grounds on which it judges the technical level to be sufficient/insufficient and (2) opinion of concerned party? | TC: 1.882<br>DP: 1.250<br>GA: 2.000 | [Example of Comment]<br>✓ Data and information are not described.<br>[Analysis]<br>✓ The average score for TC and DP is low. Analyzing the sustainability of these schemes from a technical point of view seems to be rather complicated and it is not easy to derive reasonable evaluation judgments. |

# 2. Analysis by evaluation criteria (Sustainability)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 23 | [Technical Aspect] With respect to the grounds on which the technical levels are judged to be sufficient/insufficient, (1) does the evaluation report show data in the explanation? (2) Is the judgement consistent with analysis? | | | TC: 1.912<br>DP: 1.500<br>GA: 2.000 | [Example of Comment]<br>✓ Insufficient analysis for evaluation judgment.<br>[Analysis]<br>✓ The average score for the TC and DP is low. Analyzing the sustainability of these schemes from a technical point of view seems to be rather complicated and it is not easy to derive reasonable evaluation judgments. |
| 24 | [Financial Aspect] Does the evaluation report present (1) the data regarding the comparison between the planned and actual budget, future budget, and (2) the opinion of concerned party? (3) If the report fails to present and analyze the budget data, is the financial condition not judged as "no problem," etc., without presenting supplemental information? (No it isn't => tick "Yes"; Yes it is => tick "No") | | | TC: 1.941<br>DP: 1.250<br>GA: 2.000 | [Example of Comment]<br>✓ Financial data, etc. are only shown briefly in the evaluation result sheet, and it is better to have a little more detailed explanation.<br>✓ There is no presentation of data, and insufficient information to supplement it.<br>[Analysis]<br>✓ It is understandable that there are restrictions on conducting evaluation surveys, such as difficulties in obtaining financial data, difficulties in opening the data through publication of evaluation result sheet, etc. However, in performing the evaluation, it is important how to conduct a convincing analysis under such constraints.<br>✓ The average score of DP is low. The reason is that the total number of DP is small. The scoring of evaluation which the budget information for the implementation of the planned plan was not obtained is deducted, and it affects average score. |
| 25 | [Financial Aspect] With respect to the grounds on which the financial condition is judged to be sufficient/insufficient, (1) does the evaluation report show data in the explanation? (2) Is judgement consistent with the analysis? | | | TC: 1.941<br>DP: 1.500<br>GA: 2.000 | [Example of Comment]<br>✓ Insufficient analysis for evaluation judgment.<br>✓ Evaluation judgment seems too strict.<br>[Analysis]<br>✓ The average score of DP is low. The reason is that the total number of DP is small. The scoring of evaluation which the budget information for the implementation of the planned plan was not obtained is deducted, and it affects average score. |

# 2. Analysis by evaluation criteria (Sustainability)

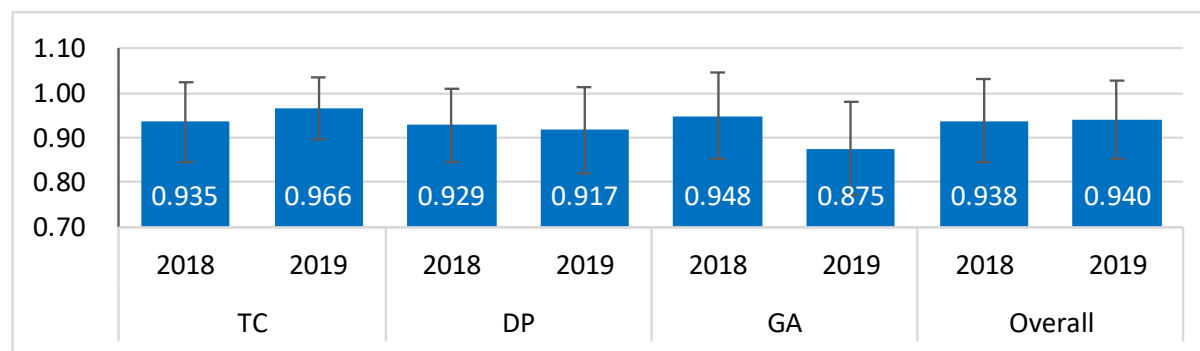| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 26 | If issues are found in individual aspects, does the evaluation report mention (1) the factors behind such issues and (2) prospects? | | | TC: 1.308<br>DP: 1.500<br>GA: 1.500 | [Example of Comment]<br>✓ Insufficient statement of prospects.<br>✓ Insufficient explanation of the factors.<br>[Analysis]<br>✓ This checklist item had a low average score overall. In many evaluations, the factors of the issues are described, but not the prospects of the issues in the future. |
| 27 | [Evaluation Result] Is the evaluation judgment on Sustainability as a whole consistent with the judgments on (1) policy aspect, (2) institutional aspect, (3) technical aspect and (4) financial aspect? | | [Evaluation Result] Is the evaluation judgment on Sustainability as a whole consistent with the judgments on (1) institutional aspect, (2) technical aspect and (3) financial aspect? | TC: 1.941<br>DP: 2.000<br>GA: 1.917 | [Example of Comment]<br>✓ Questioning in evaluation judgment (Is "XXX" seems appropriate instead of "XXX")<br>✓ Since the judgment of each aspect is not clearly described, it is difficult to derive the basis for deriving the overall evaluation judgment.<br>[Analysis]<br>✓ In general, most of the evaluations are without major problem. |
| 28 | No Item | | [Evaluation Result] (1) Does the evaluation report mention the situation about Operation and Maintenance of facilities and equipment on institutional aspect, technical aspect, or financial aspect properly? (2) Does this reflect on evaluation judgement? | TC: -<br>DP: -<br>GA: 2.000 | [Analysis]<br>✓ It is a checklist item only for GA, and in general, most of the evaluations are without major problem. |

# 2. Analysis by evaluation criteria (Conclusion, Recommendation and Lessons Learned)

**Average, dispersion, and standard deviation of "Conclusion, Recommendation, and Lessons Learned" based on standardized scores of 50 targeted project**

| | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Year | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.935 | 0.966 | 0.929 | 0.917 | 0.948 | 0.875 | 0.938 | 0.940 |
| Dispersion | 0.008 | 0.005 | 0.007 | 0.009 | 0.009 | 0.011 | 0.008 | 0.008 |
| Standard deviation | 0.090 | 0.068 | 0.082 | 0.096 | 0.097 | 0.104 | 0.092 | 0.088 |

- ✓ The average score of TC is slightly higher, and the average score of GA was slightly lower.



Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

- ✓ As for the conclusion, in general, there are little description with problem.
- ✓ As for recommendations, the average score was slightly lower. Although the necessity of the recommendations is understandable, some recommendations have little relevance to the evaluation results. Similarly, for lessons learned, the average score was slightly lower. Although the importance of lessons learned is understandable, some lessons learned were not closely related to the evaluation results.
- ✓ The scoring of third party QC is high for the evaluations without recommendations and lessons learned. However, it is difficult to judge the necessity of describing recommendations and lessons learned only by the appearance of evaluation result sheet, and as a result, the number of ticks in "yes" increased in third party QC.

# 2. Analysis by evaluation criteria (Conclusion, Recommendation and Lessons Learned)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 29 | [Summary of the Evaluation] Is the overall evaluation judgment consistent with the description of evaluation results? | | | TC: 2.000 DP: 2.000 GA: 1.833 | [Analysis]<br>✓ In general, most of the evaluations are without major problem. |
| 30 | [Recommendations] (1) Is the reason(s) why the proposed recommendation should be followed (i.e., issues, etc.) in relevant sections of evaluation report ? (2) Is the necessity of recommendation explained enough? | | | TC: 1.839 DP: 1.667 GA: 1.583 | [Example of Comment]<br>✓ The reasons for the recommendations are not explained in the evaluation results.<br>✓ The recommendation is somewhat lacking in concreteness and feasibility.<br>✓ The recommendation includes too much general content.<br>✓ The necessity is not indicated.<br>✓ It is unclear how it relates to the resettlement of the issues.<br>✓ The grounds of the recommendation are unclear.<br>✓ It has no relation with the description on the evaluation result sheet.<br>[Analysis]<br>✓ The average score is slightly lower. The third party QC comments are as described above.<br>✓ Although the necessity of the recommendations is understandable, some recommendations have little relevance to the evaluation results. It is important to conduct surveys and information collection with consideration for derivation of recommendations, and to clarify the relation between evaluation results and recommendation in the evaluation result sheet.<br>✓ It is also important to keep in mind the feasibility of the recommendations (for example, how to secure a budget if it should be secured). Focusing on improving the feasibility of recommendation, It is important to collect necessary information during evaluation surveys, and demonstrate persuasive recommendation by clearly showing the relationship between recommendation and evaluation results. |

# 2. Analysis by evaluation criteria (Conclusion, Recommendation and Lessons Learned)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 31 | [Lessons Learned] (1) Does the evaluation report mention in relevant sections the finding from which you drew the lessons learned? (2) Does this bring the useful lessons learned? | | | TC: 1.941<br>DP: 1.667<br>GA: 1.778 | [Example of Comment]<br>✓ It is too abstract.<br>✓ The necessity of lessons learned is questionable. (e.g. The need for data collection for the indicators of the overall goal during the project implementation period is a lesson learned, but the overall goal is to confirm the situation several years after project completion, not during the project implementation period.)<br>✓ The lesson learned seems that it was not derived from the evaluation results.<br>✓ The description is difficult to understand.<br>✓ The contents described in the lessons learned are mainly just factual and are not described as lessons learned to be utilized in the future.<br>✓ There seems no relation between lessons learned and the evaluation result.<br>✓ [Analysis]<br>✓ The average score is slightly low. Third party QC comments are as described above.<br>✓ Although the importance of lessons learned is understandable, some lessons learned are not closely related to the evaluation results. |
| 32 | In case of no description of recommendation or lessons learned, is such way of description adequate? (Example of inadequate description: There is significant problem in impact or sustainability, but there is no recommendation for coping with the problem; There is factor analysis that might lead recommendation, but there is no recommendation derived, etc.) | | | TC: 2.000<br>DP: 2.000<br>GA: 2.000 | [Analysis]<br>✓ The scoring of third party QC is high.<br>✓ However, it is difficult to judge the necessity of describing recommendations and lessons learned from the appearance of evaluation result sheet, and as a result, the number of ticked in "yes" increases.<br>✓ In case of evaluations that clearly require analysis of recommendations and lessons learned, such as the cases that the evaluation judgement of "effectiveness/impact" and/or "sustainability" is "low", the recommendations and lessons were clearly described as required. |

# 2. Analysis by evaluation criteria (General Matters)

**Average, dispersion, and standard deviation of General Matters based on standardized scores of 50 targeted project**

| Year | TC | | DP | | GA | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| Average | 0.924 | 0.980 | 0.893 | 1.000 | 0.917 | 1.000 | 0.918 | 0.987 |
| Dispersion | 0.024 | 0.008 | 0.033 | 0.000 | 0.020 | 0.000 | 0.024 | 0.005 |
| Standard deviation | 0.155 | 0.090 | 0.182 | 0.000 | 0.141 | 0.000 | 0.141 | 0.074 |

✓ There is no particular tendency for each scheme.

✓ In general, most of the evaluations are without major problem.
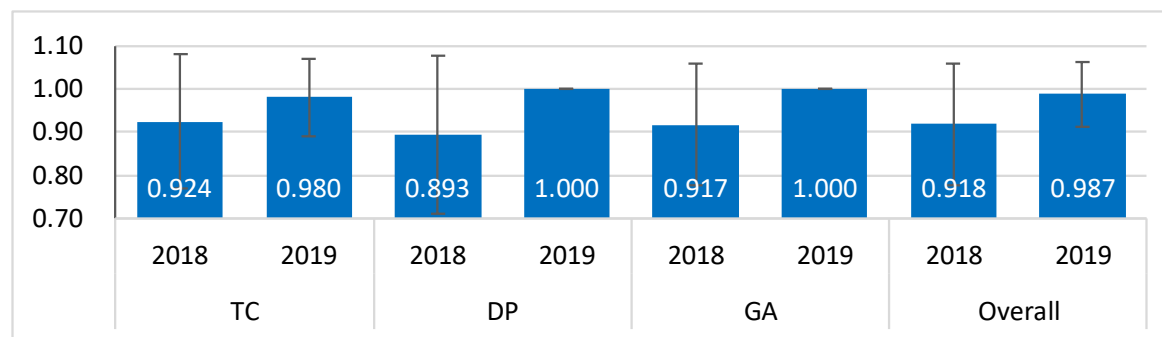


Note : Bar graph shows average of standardized score and error bar (error range) shows standard deviation.

**[Analysis of individual checklist items (for details, see the following pages)]**

✓ In general, most of the evaluations are without major problem.

✓ However, regarding the restrictions and constraints to be considered in evaluation, it is difficult to check by third party checker from the appearance of evaluation result sheet, and only the described constraints were checked. Similarly, it is difficult to determine whether all the information sources were indicated from the appearance of evaluation result sheet, and only the described information source, such as the sources of indicator data was checked).

✓ Regarding the checklist item of sample survey, the evaluations clearly uses the sample survey are checked and scored in third party QC. On the evaluation result sheet, whether or not to conduct a sample survey is not described, so the decision of third party checker depends on its content. However, in some evaluations, it was difficult to decide whether the survey was a sample survey or a mere interview with the stakeholders.

# 2. Analysis by evaluation criteria (General Matters)

| | Checklist items (TC) | Checklist items (DP) | Checklist items (GA) | Average (Raw score) | Examples of third party QC Comment and Analysis |
|---|---|---|---|---|---|
| 33 | If there are important constraints on evaluation (e.g., inappropriate indicator(s), inadequate information collection, etc.), (1) does the evaluation report explain them and (2) how they might have affected evaluation judgment? | | | TC: 2.000 DP: 2.000 GA: 2.000 | [Analysis] ✓ The scoring of the third party QC is high. However, it is difficult for the third party checker to judge whether or not there are any restrictions or constraints to be considered only from the appearance of evaluation result sheet. As a result, only the evaluations clearly indicate those constraints are the focus of third party QC and scoring tend to be a little biased. |
| 34 | Does the evaluation report identify the sources for all information it presents? ("According to the implementing agency" is acceptable) | | | TC: 2.000 DP: 2.000 GA: 2.000 | [Analysis] ✓ The scoring of third party QC is high. However, it was difficult for the third party checker to judge whether or not "the sources for all the information" are described in the evaluation result sheet. As a result, the scores tended to be higher. ✓ Regarding the indicator data, the source was described in all evaluations. |
| 35 | In case the evaluation report states it conducted a sample survey, (1) does it mention the number of valid responses and (2) the attributes of the respondents? | | | TC: 1.625 DP: All are ticked in "Not Applicable" GA: 2.000 | [Example of Comment] ✓ The number of respondents is not indicated. [Analysis] ✓ Obviously, there are few sample surveys, and the number of ticks in "not applicable" increases. ✓ It should be noted that since the implementation of sample survey is not clearly described in the evaluation result sheet, the third party checks whether or not the sample survey is conducted bases on the content of the evaluation result sheet. In many evaluations, it is difficult to judge whether it conducted a sample survey or a mere interview with stakeholders. |