Part 4

Secondary Evaluation by the Advisory Committee on Evaluation

Secondary Evaluation by the Advisory Committee on Evaluation

Seiji Kojima
Vice-President
Chairperson of JICA Evaluation Study Committee

JICA established the Advisory Committee on Evaluation in fiscal 2002 and since then has committed itself to enhancing the evaluation system and improving projects using evaluation results while receiving advice from external experts. As part of that effort, in order to increase transparency and objectively in evaluation results, the Advisory Committee on Evaluation has evaluated terminal evaluations performed by JICA (secondary evaluation) and the results have been published in the Annual Evaluation Report since fiscal 2003. In fiscal 2005, which is its third year, with the help of the Japan Evaluation Society, JICA set up a working group consisting of third-party experts in evaluation under the Advisory Committee on Evaluation to conduct secondary evaluations. Part 4 provides results of secondary evaluations conducted by the Advisory Committee on Evaluation and its working group.

As was the case last year, this year's secondary evaluation focused on examining the quality of terminal evaluation (primary evaluation). Also, secondary evaluators, being the third party, re-examined the results of the primary evaluation based on the information contained in the evaluation reports. In terms of quality of evaluation, the result of the secondary evaluation, which rates eight evaluation criteria comprising 33 viewpoints, shows that all evaluations received more than three points on a scale of five and over-the-year improvement in quality was observed when comparing the secondary evaluation results of the last three years. Nonetheless, rating on the participation of developing countries in evaluation, evaluation on efficiency, and the extraction of lessons were relatively lower than other evaluation viewpoints, leading to a conclusion that further efforts are needed. With respect to projects, after comprehensive review of the information in primary evaluation reports from the third party's perspective, re-examination was conducted based on the DAC Five Evaluation Criteria on a scale from one to five. As a result, 35 projects out of 45 were graded either "excellent" (20 points or more in the full 25 points) or "good" (15 points or more and less than 20). However, 10 other projects were graded "poor" (10 points or more and less than 15). It should be noted that severe evaluation was granted to some projects in the secondary evaluation due to an inappropriate value judgment in the primary evaluation with poor quality, though it was rated good in the primary evaluation.

With regard to the quality of evaluation, based on the results of the secondary evaluation, JICA has been working to revise JICA's Evaluation Guidelines, promote evaluation training, and share highly regarded cases of secondary evaluation within the organization, since the introduction of secondary evaluation. Moreover, using the secondary evaluation check sheet, JICA staff has controlled the quality of primary evaluation. We are very delighted with the commendation that the quality of evaluation has steadily improved as a result of these efforts, although many issues remain unsolved. We are determined to continue to improve the quality of evaluations based on the results of the secondary evaluation.

As for project evaluation, since the environment surrounding a project differs depending on the target country and sector, it is difficult to compare them in a uniform manner. Some of the projects rated "poor" should have been planned more meticulously and should have sought more appropriate project management at the implementing stage, including the review of the plan in line with changes in circumstances. At the same time, there were cases where financial sustainability was threatened by economic hardship in the target country and cases where cooperation activities were greatly curtailed due to political and security reasons in the target country. In the latter cases, the efforts made by stakeholders in these projects are not at all inferior to the efforts made by the stakeholders engaged in the projects which have been rated "excellent." Still, being resolutely results-oriented as an implementing body of government-funded ODA projects, we must solemnly accept the fact that there are projects that have been rated "poor" in the view of the secondary evaluators who have expertise in ODA and evaluation.

By having evaluation results re-examined from the viewpoints of the third party, JICA will take further steps to review its own projects and implement more effective and efficient projects. As an extension of that effort, JICA will select some projects based on the results of secondary evaluation, and have some members of the Advisory Committee on Evaluation to conduct field studies to present recommendations, including the results of the field studies, for the improvement of the quality of evaluation and implementation of effective and efficient projects.

Last but not least, I would like to express my sincere gratitude to every member of the Advisory Committee on Evaluation and its working group for offering valuable comments. All the members carefully examined 45 terminal evaluation reports from various aspects and performed secondary evaluation by devising various measures that enable more reliable and convincing secondary evaluation, despite the constraints of the nature of the secondary evaluation.

Chapter 1 Results of Secondary Evaluation Fiscal 2005

Advisory Committee on Evaluation/
Secondary Evaluation Working Group

1-1 Objectives, Targets, Methods of Evaluation

(1) Objectives

A number of options are possible as to who shall conduct evaluation on ODA projects. If evaluations are performed by stakeholders, it is expected that detailed evaluation in light of circumstances are possible since the evaluators have profound knowledge of the project and region and fully understand the activities and various situations. Also, feedback will more likely fully function, leading to improvement in the project. On the other hand, it could result in lenient evaluations since they may make too much allowance for circumstances, which gives rise to problems in transparency and neutrality. Due partly to the nature of its operation, JICA manages a number of relatively small projects, and therefore, JICA, in reality, does not have any other choice but to conduct internal evaluation, or if not that, it has to seek the assistance of outside stakeholders, such as domestic support committee members, to conduct the evaluation. For terminal evaluation alone, the number goes to around 50 every year.

Accordingly, as a means of overcoming the expected disadvantages while taking advantage of internal evaluation, objectivity and neutrality can be achieved by conducting internal evaluation thoroughly in compliance with the guidelines and through secondary evaluation by external experts on the results of the internal evaluation. In other words, in order to evaluate a number of projects, it is practical to develop a system where the results of internal evaluation are reviewed and authorized if the results are good, and modified if not.

The introduction of the Plan-Do-Check-Act (PDCA) cycle is effective in constantly improving projects. Evaluation corresponds to the Check part of this cycle. If this concept is applied to evaluation, the PDCA cycle of evaluation (planning of evaluation - implementation of evaluation - evaluation of evaluations - improvement of evaluation) becomes complete. In order to avoid partial and subjective evaluation, it is important to incorporate the views of external examiners; however, they do not necessarily have to evaluate every single project. At least, a certain level of transparency and objectivity can be secured if the view of the external examiners is incorporated into the Check part of the PDCA cycle.

Evaluation is a set of processes, from collecting information and performing analysis/evaluation to drawing out recommendations/lessons and compiling reports in an evaluation framework. In order to ensure reliability of primary evaluation such as terminal evaluation as in the previous years and facilitate the disclosure of easy-to-understand evaluation results, the secondary evaluation in fiscal 2005 was performed with a focus on the following questions.

a. Evaluation of the quality of primary evaluation
- Does the primary evaluation satisfy a certain quality?
- Has the quality of primary evaluations improved every year?
- What tasks should be carried out to further upgrade the quality?

b. Evaluation of projects by secondary evaluators based on the reports of primary evaluation
- What is the result of secondary evaluation of the project?
- Is there any relation between the results of secondary evaluation on the project and the quality of primary evaluation?

(2) Evaluators

Now, there is a question about who conducts secondary evaluation. It is better to perceive that the value of secondary evaluation is determined by whether the evaluation results themselves are convincing, rather than whether they are correct or incorrect. There is no single answer to the question of how evaluation should be carried out, but the answer varies depending on the evaluator's background and the sense of value that affects the evaluation. If numerical targets are set for projects, there may be less chance of disagreement over whether the project purposes have been achieved or not. It is still natural that there are differences in opinions on the reasons and response measures. Even if a secondary evaluator has been provided, there is no guarantee that her secondary evaluation result is the utmost and foremost. It is quite probable that results are different when another evaluator conducts secondary evaluation. If so, it is safer and more practical to come up with a framework to allow opinions of several secondary evaluators with some level of ability, rather than finding one excellent evaluator.

JICA has established the Advisory Committee on Evaluation to solicit opinions on the nature of JICA evaluation and evaluation results. However, due to the nature of the committee, the opinions are inevitably general, making it difficult to conduct detailed secondary evaluation on each evaluation result. Thus, it is practical to set up a working group to perform secondary evaluation, take time to examine each of the internal evaluation results and further discuss the outcomes at the parent committee meetings.

It was in fiscal 2003 that JICA launched the secondary evaluation on the terminal evaluation by setting up the Secondary

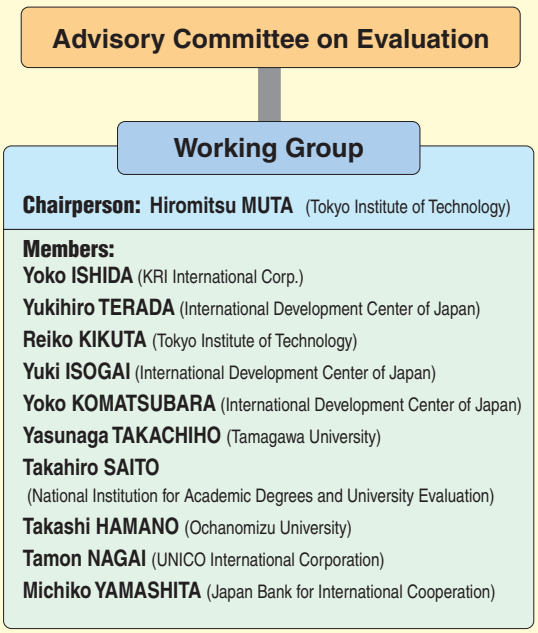
Evaluation Working Group. Eight members of the Advisory Committee on Evaluation took charge of the secondary evaluation in fiscal 2003. In fiscal 2004, the Secondary Evaluation Working Group was formed under the Advisory Committee on Evaluation, consisting of six experts and eight JICA staff members (primarily evaluation chiefs of each department). The experts were selected on the basis of recommendations by the Japan Evaluation Society to guarantee the objectivity of the selection. The differences in evaluation tendency between experts and JICA staff were also explored. As a result, no significant difference in evaluation tendency between them was confirmed.

As the development of methodology is almost complete thanks to the practice in fiscal 2004, the prospects for practical application have become bright. Taking advantage of the expert knowledge of external intellectuals, the work of the secondary evaluation was subcontracted to the Japan Evaluation Society in fiscal 2005. The Japan Evaluation Society formed an evaluation team comprising 10 members. The members were recruited within the Society in an attempt to increase transparency. The secondary evaluation system of fiscal 2005 is shown in Figure 4-1.

(3) Evaluation Targets

This secondary evaluation targeted 28 terminal evaluations conducted in fiscal 2003 and 17 terminal evaluations in fiscal 2004. The main targets of this fiscal year's analysis are the total of these 45 projects. For a year-to-year comparison purpose, an additional 10 terminal evaluations conducted in fiscal 2003 and 11 evaluations in fiscal 2002, which had been subject to the previous secondary evaluation, were sampled without bias and targeted under the secondary evaluation of this year (Appendix 1).

Figure 4-1 Secondary Evaluation System



(4) Evaluation Design and Methods

If every member of the secondary evaluation reads and evaluates all the evaluation reports, the mean scores for each evaluation criterion (evaluation viewpoint/criteria) will reflect the opinions of all the evaluators and the results will be free of personal evaluation bias. This is because the results will be biased unless the opinions of a certain number of evaluators are averaged since each evaluator has different backgrounds and opinions. However, this method is impractical due to the enormous workload placed on each evaluator. In fact, it takes two to three hours for an evaluator to thoroughly read an evaluation report and fill in the scores and comments on an evaluation sheet. When overlapping evaluation reports from the previous years are included, the number of reports subject to the secondary evaluation exceeds 60 per year. It takes an enormous amount of time and effort to read and evaluate all those reports, although it is not impossible.

Thus, it is effective to split the work and in fiscal 2005, each evaluation report was read by four evaluators under the appropriate assignment to avoid bias. Specifically, two key members of the evaluation working group read all the reports, one member read 27 reports, and seven members read 15 to conduct the secondary evaluation. This scheme allows us to treat the judgment criteria of the two key members as the norm of the entire group and adjust the judgment criteria of the other evaluators. Although fairness is more likely if the results represent the average of four evaluators, rather than one, it is still unavoidable that the tendency of the evaluators could affect the results since each report is rated by only four evaluators, raising the question of credibility and impartiality of the secondary evaluation. Theoretically speaking, the scores given by each evaluator for each evaluation criterion can be divided into two parts: true score of the evaluation target (free of personal evaluation bias of the evaluator) and coefficient of evaluation tendency of an individual evaluator (strictness/leniency coefficient: error tendency of individual evaluator). Accordingly, a method of statistical analysis (analysis of variance) was devised to remove those two parts and the evaluation tendency of evaluators was adjusted to obtain the unbiased estimate of evaluations scores that are free of personal evaluation tendency.

A year-to-year comparison was made possible through appropriate sampling by repeating the evaluation of the same project for several years. The projects that had been evaluated repeatedly for two fiscal years can serve as so-called "overlap width" for unification. Using the overlap width, it is possible to link the secondary evaluation results of fiscal 2004 with those of fiscal 2005. True estimates of the evaluation scores were calculated for fiscal 2004 and fiscal 2005; however, the evaluation standards seem to be different. Since the objective was to see the distribution of evaluation scores, it was first necessary to think of a way to

match the mean scores and variances of both fiscal years in the overlap width. Then, the results of the secondary evaluation in fiscal 2004 were converted in such a way so that the mean scores and variances in the overlap width for each fiscal year corresponded to each other. Since the sampling in the overlap width was conducted appropriately, such a simple conversion was sufficient to make a comparison. In this way, evaluation data obtained individually can be processed and analyzed as a large pooled sample through the unification of disconnected evaluation information in various ways using the overlapping evaluations (overlap width).

(5) Evaluation Methods

As mentioned earlier, the secondary evaluation of terminal evaluation has two objectives. One is to evaluate the quality of terminal evaluation and the other is to validate the evaluation results of terminal evaluation. Basically, in a secondary evaluation several experts evaluate the evaluation results (reports) based on a set of evaluation viewpoints. It was in fiscal 2003 that JICA started secondary evaluation by experts on terminal evaluation, and the secondary evaluation was conducted on 40 projects at the time. First, the secondary evaluators read the terminal evaluation reports and evaluated them on a five-point scale using the designated secondary evaluation check sheet. Evaluations were made from the perspective of the quality of the terminal evaluation (eight criteria with respect to evaluation framework, six criteria with respect to implementation of studies, nine criteria with respect to information analysis/evaluation, and four criteria with respect to recommendations/lessons: a total of 27 criteria) as well as the perspective of evaluation of the project itself (six criteria). Evaluation items listed in the evaluation sheet and evaluation criteria were made based on the criteria for good evaluations in the JICA Evaluation Guidelines.

In fiscal 2004, using the evaluation analysis of the previous year as a reference, the burden of the evaluators was reduced and the evaluation viewpoints were improved to structure the evaluation criteria. On the sheet, evaluation scores were given to criteria on a scale of 10 and several viewpoints shown for consideration of each grading were evaluated on a scale of three. In addition to 58 projects for fiscal 2003-2004, 10 overlapping projects were evaluated again for the purpose of a comparison with the previous year. In fiscal 2005, in addition to 45 projects for fiscal 2003-2004, 21 overlapping projects for fiscal 2002-2003, on which the secondary evaluation was conducted in the previous year, were evaluated again for the purpose of a year-to-year comparison.

Evaluators are not the same members every year. Even if they are, after the interval of one year, there is no guarantee that the person will evaluate on the basis of the same evaluation standards. In order to see the over-the-year changes in the evaluation results, an adjustment of evaluation standards of evaluators is necessary. While referring to the evaluation results of the previous year, 21 projects of fiscal 2002-2003 were carefully sampled with consideration of balance out of the projects on which the secondary evaluation had been performed in the previous fiscal year, in order to perform a comparison with the previous year.

Further improvements were made on the evaluation criteria this year. The comparison between the evaluation viewpoints between fiscal 2004 and fiscal 2005 is shown in Table 4-1. Together with the reorganization and integration of evaluation viewpoints, both ratings for viewpoints and criteria were changed to a scale of five. This is because the psychological burden on the evaluators can be reduced if the scales are uniform. As you can see, the evaluation viewpoints of fiscal 2004 and 2005 are by and large similar overall, allowing year-to-year comparison.

Table 4-1 Comparison of Evaluation Viewpoints between Fiscal 2004 and 2005

Fiscal Year	2004		2005	
Evaluation Criteria	Viewpoints	Rating	Viewpoints	Rating
Evaluability	4 (3-point scale)	10-point scale	4 (5-point scale)	5-point scale
Evaluation Framework	4 (3-point scale)	10-point scale	3 (5-point scale)	5-point scale
Data Collection	5 (3-point scale)	10-point scale	4 (5-point scale)	5-point scale
Assessment of Performance	4 (3-point scale)	10-point scale	4 (5-point scale)	5-point scale
Analysis	3 (3-point scale)	10-point scale	3 (5-point scale)	5-point scale
Evaluation	7 (3-point scale)	10-point scale	6 (5-point scale)	5-point scale
Recommendations/Lessons Learned	8 (3-point scale)	10-point scale	6 (5-point scale)	5-point scale
Reporting	4 (3-point scale)	10-point scale	3 (5-point scale)	5-point scale
General Criteria for Good Evaluation	4 (3-point scale)	10-point scale		
Evaluation of the Project: Relevance		10-point scale	3 (5-point scale)	5-point scale
Evaluation of the Project: Effectiveness		10-point scale	2 (5-point scale)	5-point scale
Evaluation of the Project: Efficiency		10-point scale	2 (5-point scale)	5-point scale
Evaluation of the Project: Impact		10-point scale	3 (5-point scale)	5-point scale
Evaluation of the Project: Sustainability		10-point scale	5 (5-point scale)	5-point scale
Evaluation of the Project: Overall Evaluation		10-point scale		

Table 4-2 Secondary Evaluation Criteria

I.	Criterion:	The precondition for conducting appropriate evaluation was possible (Evaluability)
	Viewpoints:	Evaluability of the Initially prepared Project Design Matrix (PDM)
		Evaluability of Outputs, Project Purpose and Overall Goal
		Logic of Project Design
		Project Monitoring
II.	**Key Evaluation Criteria**	
1.	Criterion:	Evaluation Framework
	Viewpoints:	Time Frame of Evaluation Study
		Evaluation Team Composition—Impartiality and Specialty
		Level of Counterpart Participation
2.	Criterion:	Data Collection
	Viewpoints:	Evaluation Questions
		Appropriateness of Data Collection Methods and Data Sources
		Data/Information Sources
		Sufficiency of Data/Information Obtained
3.	Analysis/Evaluation	
3.1	Criterion:	Assessment of Performance
	Viewpoints:	Measurement of Results
		Examination of Project Implementation Process
		Examination of Causal Relationships —Logic of Project Design
		Examination of Causal Relationships —Before and After
3.2	Criterion:	Analysis
	Viewpoints:	Objectivity of Analysis
		Holistic Analysis
		Analysis of Promoting and Impeding Factors
3.3	Criterion:	Five Evaluation Criteria
	Viewpoints:	Relevance
		Effectiveness
		Efficiency
		Impact
		Sustainability
		Conclusion
4.	Criterion:	Recommendations/ Lessons Learned
	Viewpoints:	Relevance and Credibility of Recommendations
		Sufficiency of Recommendations
		Usability of Recommendations
		Relevance and Credibility of Lessons Learned
		Sufficiency of Lessons Learned
		Usability of Lessons Learned
5.	Criterion:	Reporting
	Viewpoints:	Presentation/Legibility and Clarity
		Utilization of Tables and Figures
		Presentation of Primary Data
III.	**Evaluation of the Project Based on the Report**	
1.	Criterion:	Relevance
	Viewpoints:	Validity
		Necessity
		Appropriate Approach
2.	Criterion:	Effectiveness
	Viewpoints:	Achievement Level of Project Purpose
		Causal Relationships between Outputs and Project Purpose
3.	Criterion:	Efficiency
	Viewpoints:	Cost-effectiveness
		Appropriate Implementation Process
4.	Criterion:	Impact
	Viewpoints:	Achievement Level of Overall Goal
		Causal Relationships between Project Purpose and Overall Goal
		Unintended Positive and Negative Impact
5.	Criterion:	Sustainability
	Viewpoints:	Mechanism of Securing Sustainability
		Level of Sustainability
		Organizational Sustainability
		Technological Sustainability
		Financial Sustainability

The evaluation viewpoints in fiscal 2005 are shown in Table 4-2 and Appendix 2. Analysis was made based on these evaluation viewpoints. Evaluation was made on the basis of the following five-point scale for rating both viewpoints and criteria.

5: Sufficient/high
4: Fairly sufficient/high
3: Average
2: Slightly insufficient/low
1: Insufficient/low

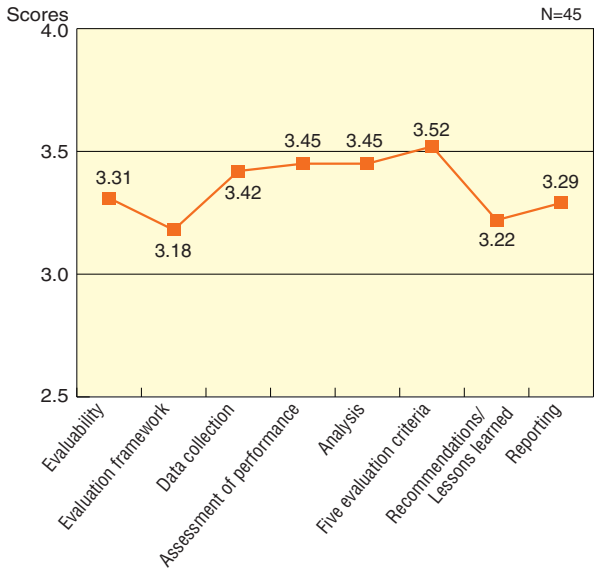
1-2 Quality and Challenges of Terminal Evaluation Examined through Reports

(1) Overview of Evaluation Results

The average scores for individual evaluation criteria of the 45 terminal evaluations conducted in fiscal 2003 and 2004 are shown in Figure 4-2. All the average scores are over 3.0 points and belong to the level of "average" or higher in the grading scale. The scores are relatively high for the criteria of "data collection" for evaluation, "assessment of performance" in analysis, and "analysis" methods and "five evaluation criteria" of DAC's five evaluation criteria; however, the average scores for "evaluation framework" concerning time frame and composition of study team and "recommendations/lessons learned" are slightly lower.

When looking at the distribution of scores by evaluation criterion, as shown in Figure 4-3, many are distributed between 2.5 and 4.49 and in particular the scores between 3.0 and 3.99 constitute a high proportion. However, the distribution exhibits different patterns depending on the evaluation criterion. The scores of "evaluability" range from 1.5 and 4.49 points and those for "five evaluation criteria" between 2.0 and 5.0, indicating a large variation of the quality of terminal evaluation. On the other hand,

Figure 4-2 Average Scores by Evaluation Criterion



"assessment of performance" ranges between 2.5 and 4.49 and "recommendations/lessons learned" between 2.0 and 3.99, with little variance. Most of the scores for "evaluation framework" are in the range of 3.0 to 3.49, with a few scores above 3.5, showing uneven distribution.

In summary, with regard to the quality of terminal evaluations of the target projects, many projects were given 3.0 points ("average") or higher and some were given 4 points ("good"); all the scores in evaluation criteria achieve a certain quality of "average" on average. However, few projects were given 4.0 points or higher in the criteria of "evaluation framework" and "recommendations/lessons learned," leading to lower average scores than other evaluation criteria.

(2) Evaluation Results by Criterion and Issues for the Improvement of Quality

In the secondary evaluation, the viewpoints of each evaluation criterion were rated, and qualitative evaluation information was collected in the form of comments of the evaluators that were written in the additional box on the sheet. We will discuss the current conditions and issues of the quality of terminal evaluation by criterion based on the evaluation results of scores for the viewpoints of each evaluation criterion and the comments from the evaluators. Figure 4-4 shows the average scores for viewpoints under each evaluation criterion as well as those for evaluation criteria.

a. Evaluability

Evaluability is a criterion to see whether an appropriate evaluation was possible or not. The average scores of the viewpoints under this criterion fall near 3.3 points, securing the "average" level, and they are not particularly high or low compared with those of the viewpoints under other evaluation criteria.

"Evaluability of the initially prepared project design matrix (PDM)" is a viewpoint to validate whether the initially prepared PDM was used for evaluation without much alteration, and whether the PDM used for evaluation was not drastically different from the PDM formulated at the planning stage (whether the project itself had to be drastically modified because of the changes in the project purpose and indicators of the project). Although the score related to PDM itself is satisfactory on average, there were cases where the details of project purpose and indicators did not agree with the partner country at the time of the launch of the project and where the PDM was not formulated by the time of

Figure 4-3 Distribution of Scores by Evaluation Criterion

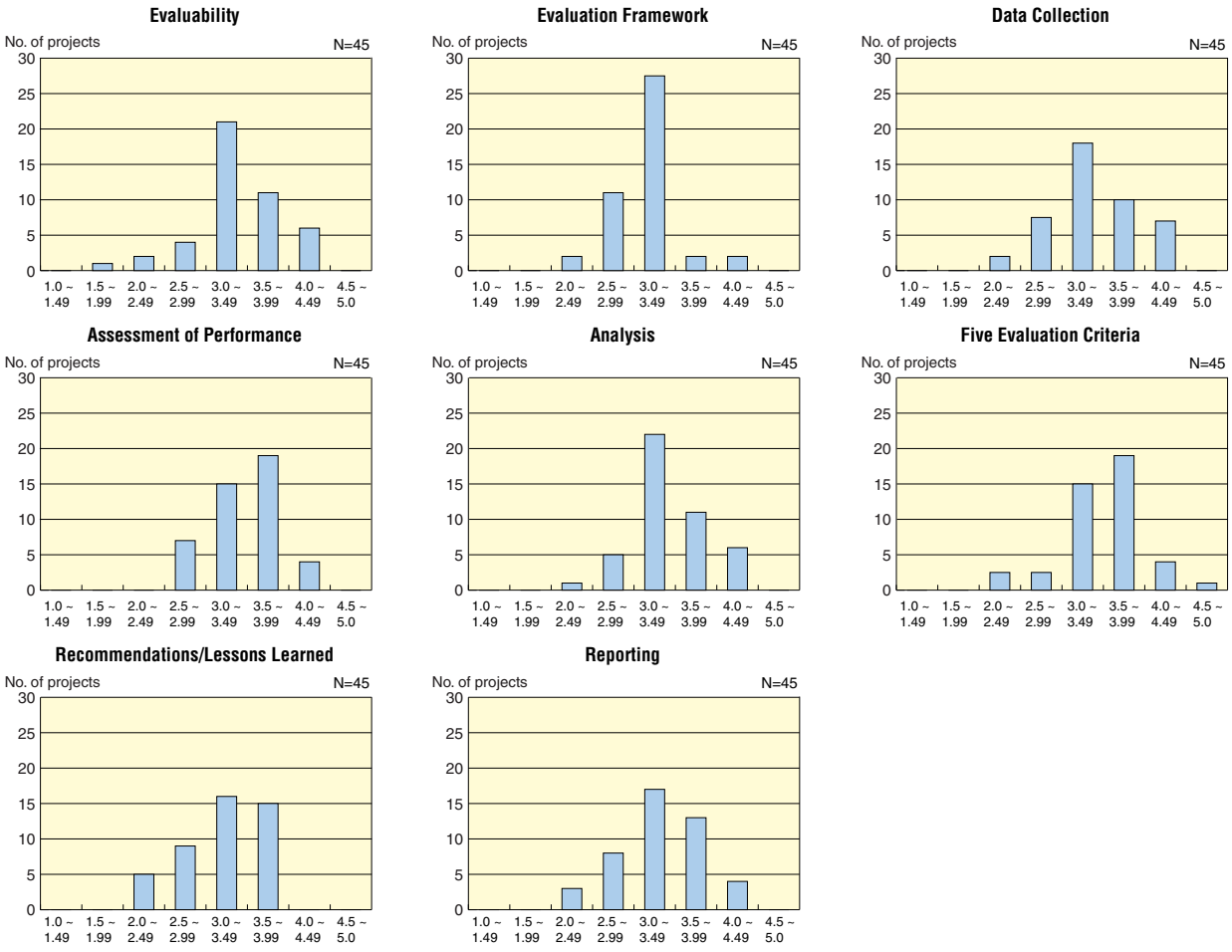
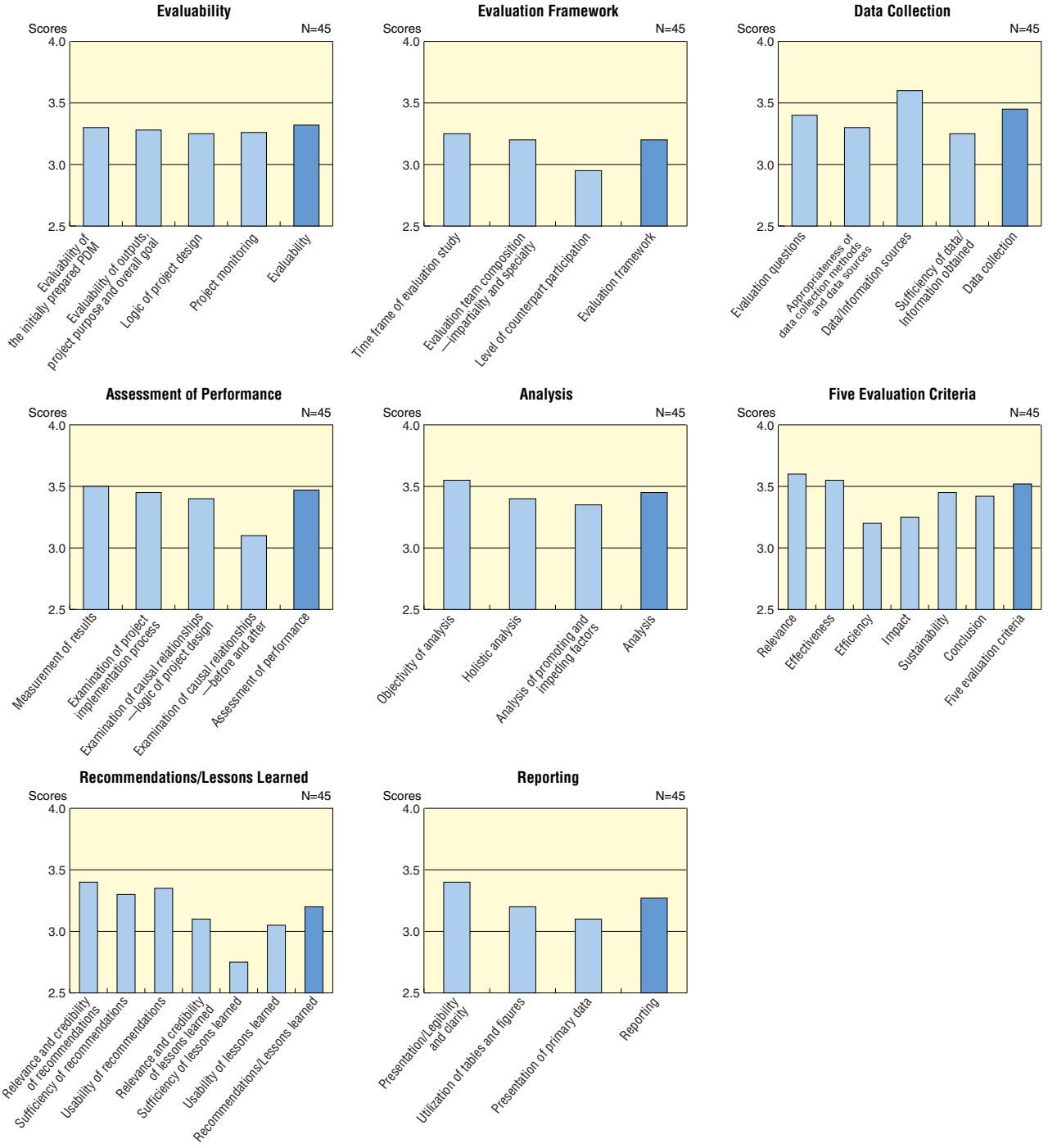


Figure 4-4 Average Scores for Viewpoints under Each Evaluation Criterion



terminal evaluation.

One project obtained high scores for "analysis" and "effectiveness" because the PDM was fully understood by all the counterparts, used as monitoring sheet for project activities, and helped in terminal evaluation. On the other hand, there is a project where the PDM was inadequately formulated; for example, indicators to measure the degree of achievement were not appropriate; there were discrepancies between the overall goal and project purpose; and logical flow from activities to the purpose was weak. It is important to set appropriate goal and purpose and develop a logical PDM in order to confirm the degree of achievement and improve the quality of evaluation.

b. Evaluation Framework

The average scores for the viewpoints of "time frame of evaluation study" and "evaluation team composition" under the criterion of evaluation framework are around 3.2, which is above the "average" level. However, the average score for "level of counterpart participation" failed to reach 3.0 points, which is lower than the scores of the viewpoints under other evaluation criteria.

As far as time frame of evaluation study is concerned, some evaluations spent only two to three days or less on data collection

according to the information from the evaluation report, which seems insufficient. In contract, there was a project where evaluation was conducted within an efficient time frame by distributing questionnaires in advance and having the advance team collect basic data from the questionnaires. In a project related to education, coinciding with the terminal evaluation, the final exam of the inaugural class took place. If the evaluation had been postponed to a later date, the exam results would have been made the indicators of outcomes and been effective to measure achievement of the project purpose. In such a case, in order to conduct useful evaluation, it is necessary to consider the timing of evaluation carefully.

Since terminal evaluation is internal evaluation, the evaluation team is more likely to consist of people concerned with the project implementing organization or Japan's supporting organizations. The participation of the supporting organizations can be an advantage in terms of specialty. In any case, since they are involved with the project, it is inevitable that neutrality may suffer. Although the report lists the organizations to which the evaluation team members belong, that alone does not specify profiles of their operations and specialties. Therefore, it is desirable to include not only the names of affiliations, but the areas of specialty as well as the relationships with the project.

The participation of the partner county in evaluation is essential not only for securing the feedback of evaluation results and capacity building of the partner country, but also for ensuring neutrality, impartiality and specialty of evaluation. Some projects succeeded in gaining sufficient participation of the partner countries by involving persons of the partner countries who are assumed to be in neutral positions or by holding comprehensive workshops including the counterpart. On the other hand, many projects are hard to judge in terms of how much the partner countries were involved in evaluation: specifically, whether evaluation was carried out jointly, or whether stakeholders of the partner countries were simply informed about the evaluation results. This resulted in the low score for the level of counterpart participation.

In terms of the level of counterpart participation, how much they are involved in designing the evaluation is critical. Evaluations used to be designed by Japanese consultants and presented to the counterpart to be agreed upon at the meeting. With more authority delegated to overseas offices, an increasing number of evaluations are designed jointly by overseas offices and partner countries. Although many reports refer to organizing a joint evaluation committee and both parties sign and exchange the joint statement of the evaluation results, it is still difficult to gauge from the reports how much or little the partner countries were involved in a series of evaluation processes, including evaluation design. In order to increase the participation of the partner countries and ensure neutrality and specialty at the same time, the report needs to clarify evaluators' specific relations with the projects and evaluation methods.

c. Data Collection

In the criterion of data collection, many projects collected data sufficiently and appropriately from a wide range of data sources thanks to a quite detailed and appropriate evaluation grid. In general, each viewpoint in data collection attained the level of above "average." The highest score is 3.6 points for "data/information sources." This is higher than the viewpoints in other evaluation criteria. Though "sufficiency of data/information obtained" scored nearly 3.3 on average and attained the level of above "average," it is still slightly lower than other viewpoints in the same criterion.

The viewpoint of "data/information sources" questions whether the sources of data/information (the locations of visits and identity of interviewees) were clarified and whether sufficient explanation concerning data sources (list of data sources and interviewees) was provided, which shows the objectivity of evaluation. The sources were shown appropriately as a whole, leading to high scores; however, some projects did not list the places of visits, interviewees, or data sources.

On the other hand, "sufficiency of data/information obtained" scored relatively low. This viewpoint questions whether the information collected was sufficient to conduct evaluation, whether sufficient data was gained to answer the evaluation questions based on the predetermined plan of information collection (evaluation grid), and whether necessary additional information was gathered for newly confronted questions during the evaluation process. Some projects with insufficient information were observed; for example, the interviewees were limited to the responsible parties of the counterpart and data collection from beneficiaries was required. Furthermore, several projects did not have or attach evaluation grids.

Evaluation questions should encompass not only indicators but also information necessary for comprehensive evaluation and should be established in such a way as to enable qualitative assessment of project purpose rather than be confined to quantitative data collection. Data should be collected to allow comprehensive and holistic evaluation in which alternative data sources are considered and used in case designated information becomes unavailable due to unexpected reasons.

d. Assessment of Performance

The average score of "measurement of results" is the highest at 3.5 points among the viewpoints in assessment of performance, and other two viewpoints, "examination of project implementation process," and "examination of causal relationships—logic of project design," exceed the level of "average." However, the average score of "examination of causal relationships—before and after" is 3.1 points and lower than other viewpoints here, though it is above the level of "average."

With respect to assessment of performance, one project used the activity progress sheet that lists goals to be achieved, imple-

mentation process and performance corresponding to the contents of activities, enabling the full understanding of causal relationships and clarifying reasons for delay in activities. Another project was evaluated with due account of influencing factors other than the project itself, such as impact of other projects, allowing sufficient assessment of performance and resulting in high evaluation scores. A low score was given to the viewpoint of "examination of causal relationships—before and after." This viewpoint questions whether the causal relationships were thoroughly examined to verify that effects for the beneficiaries have resulted from the project interventions through comparison of with/without and before/after. It was not rated high in the secondary evaluation of fiscal 2003 either. There were some projects that did not fully assess the contribution of projects. For example, although a comparison was made between before and after, analysis was not carried out to verify that the effects were brought about by the project implementation. Another example is that no data to support why it was concluded that improvements were made was provided. Many did not list the changes in important external assumptions.

e. Analysis

While "assessment of performance" evaluates the degree of assessment of facts, "analysis" evaluates the diversity and appropriateness of the methods. In other words, it is a viewpoint to see whether quantitative and qualitative analyses were both used and verification was properly performed.

All the viewpoints in analysis obtained above the "average" level, with the highest score of 3.5 points for "objectivity of analysis." The score of "analysis of promoting and impeding factors" was 3.3 points, which is lower than other viewpoints here. Nonetheless, it is still rather high compared with the viewpoints in other evaluation criteria.

Many projects adopted the evaluation analysis methods that are stipulated in the new evaluation guidelines of JICA and others, producing a convincing analysis. For example, one project made analysis with a combination of quantitative and qualitative data from a socioeconomic perspective. Other projects conducted analysis of impeding factors or objective analysis based on quantitative indicators. Another project performed comparisons with other countries and over-the-year comparisons. In addition, various forms of analyses were devised; for instance, quality information based on interviews and meetings were combined with numerical data such as the amount and timing of input by each analysis object; the obtained data were quantified as much as possible by rating the questionnaire results as a whole.

On the other hand, there were projects that did not clarify how the information on assessment of performance was analyzed to reach the conclusion, or did not carry out sufficient analysis of the factors which affected the outcome of the project. Some projects lacked objectivity and diversity because most of the analyses were based on the information from project stakeholders. Others lacked diversity since the collected data were not effectively used for analysis.

f. Five Evaluation Criteria

The scores for the "five evaluation criteria" are high in general. The highest score among the viewpoints was "relevance" with the average of 3.6 points, and the average scores for "effectiveness" and "sustainability" were both high with 3.4 points or higher. Though the scores for "efficiency" and "impact" were both above the level of "average," they are slightly low with the score for "efficiency" below 3.2 points.

"Efficiency" was rated the lowest in the secondary evaluation for fiscal 2003 as well. This viewpoint questions whether perspectives (comparison with other similar projects through cost analysis, cost-effectiveness analysis, etc.) are sufficiently covered. Many projects were not evaluated sufficiently from the perspective of cost-effectiveness, thus leading to the low score. In fact, many reports did not contain sufficient information concerning cost performance of projects. Without information about cost performance, it is impossible to perform cost accounting, as done in the private sector, and difficult to compare how much input is needed to calculate output. However, some projects are worth high praise for having considered the estimate of appropriate cost burden and the estimates of overall operating expense and benefits of project, together with appropriate input.

"Relevance" is a viewpoint that questions whether perspectives (validity and necessity of a project in light of needs of beneficiaries, consistency of policies, project implementation as an appropriate approach to problem solving, etc.) are sufficiently covered. Many projects summarized "relevance" well, but its evaluation was questionable in some projects. For example, the relevance of the support in the area was examined, but the relevance of long-term support for the same implementing organization and considerable input accompanied was not evaluated. The relevance of some projects was evaluated high even if logic was irrelevant to preconditions or overall goals. There was no mention about the relevance concerning urgency and importance of implementing the project.

The viewpoint of "effectiveness" verifies whether perspectives (achievement level of project purpose, causal relationships between outputs and project purpose, etc.) are sufficiently covered. Some projects evaluated the prospect that project purposes would be achieved by the end of the project even if no outcomes were generated at the time of terminal evaluation. If this conclusion has to be drawn, it is necessary to present the basis for the judgment of its effectiveness; otherwise, it is only be wishful thinking.

g. Recommendations/Lessons Learned

The average scores for the viewpoints concerning recom-

mendations, "relevance and credibility," "sufficiency" and "usability" are all above 3.3 points. The highest is "relevance and credibility" with 3.4 points. However, the average scores for the viewpoints concerning lessons learned are low in general. In particular, the average score for "sufficiency" is 2.8 points and this is the lowest among the viewpoints of all evaluation criteria. Lessons learned are key to the feedback of learning effect, which is one of the objectives of evaluation. The low scores for lessons learned mean that evaluation results may not be fully utilized.

Some projects provided specific, clear, and usable recommendations/lessons learned. Others did not fully reflect evaluation results, which clarified what to do for future improvement, or failed to incorporate what was mentioned in other chapters of the reports into recommendations/lessons learned. Also, some recommendations and lessons learned that could not be derived from evaluation results or lacked a clear basis were observed. As far as contents of recommendations are concerned, many of them were too general and had no specific references as to when and how to respond. Not many lessons learned were presented and they were too general in terms of content to actually serve as lessons. These points were the reasons for poor scores.

One project presented useful recommendations/lessons learned in the statement of the evaluation team leader, instead of in the section of recommendations/lessons learned. They must have been released as the statement of the leader since the agreement had not been reached with the partner country. Given that recommendations/lessons learned are useful for implementation of future similar projects, this may be a way to go about it when something is worth mentioning although it has yet to be agreed upon with the partner countries. However, further elaboration would be necessary as to the presentation.

h. Reporting

The overall rating of reporting is low. The average scores for "utilization of tables and figures" and "presentation of primary data" were below 3.3 points. Nonetheless, they were all above the level of "average."

Some reports are understandable as they were written simply and clearly using tables and figures to show basic data in the text or presenting project purposes and activities at the beginning of the text to make a flow of argument clear. On the other hand, other reports were too verbose and inconsistent with the argument to grasp the overall picture or so poorly structured that the readers needed to look in the appendix for the results. Quite a number of reports did not contain the primary data, such as the results of hearing or questionnaire surveys, which provide essential information to determine the achievement of goals. Since the objective of reporting is to clarify the results of project implementation, it is desirable to make it more understandable and readable.

(3) Examples of Good Quality Evaluation Reports and Poor Quality Evaluation Reports

The JICA Guidelines for Project Evaluation (March 2004) explains in detail important points to be considered for appropriate evaluation using specific cases with regard to key criteria such as evaluation framework, data collection, assessment of performance, analysis, five evaluation criteria, recommendations/lessons learned, and reporting. However, it is not easy to write a report that is easy to understand and high in quality. If some high quality reports of terminal evaluations are presented, these reports can serve as role models and help quality of other reports improve.

The quality of terminal evaluations was ranked according to the total scores for eight evaluation criteria: evaluability, evaluation framework, data collection, assessment of performance, analysis, five evaluation criteria, recommendations/lessons learned and reporting. The full score for each criterion is 5 points and the lowest is 1, thus making the possible total score 40 and the minimum 8. The "average" level is set to be 24 points. The top four evaluations and the worst four evaluations were selected with consideration given to the distribution of the total scores. None of the worst four evaluations achieved a total score of 24 points, being below the "average" level. The total scores for eight criteria of these eight evaluations are shown individually in Figure 4-5 and Figure 4-6. Table 4-3 shows the average scores and the differences in average scores for evaluation criteria of the top four and the worst four projects.

As clearly observed from Figure 4-5 and Table 4-3, the average scores for evaluation criteria of the top four projects are about 3.9 for "evaluability" and above 4.0 for "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting." In other words, the logical framework for setting goals was clear, data collection was appropriate and sufficient, the implementation process, performance, and effects of projects were fully examined, collected data were objectively analyzed from various aspects, and promoting and impeding factors to the onset of effects were analyzed. Furthermore, in evaluating five evaluation criteria, necessary points to be considered were covered and the reports were also clearly presented. All of these led to high ratings on the quality of terminal evaluations. On the other hand, the terminal evaluations with lowest total scores, contrary to the top evaluations, tend to have low scores on "evaluability," "data collection," "analysis," "five evaluation criteria," and "reporting."

In every evaluation criterion, the difference in the average scores between the top and worst four projects are statistically significant, and the top four projects are rated higher than the worst four. A large difference was found between the two groups in the average scores for "evaluability," "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting," indicating that these criteria are important factors

Figure 4-5 Top Four Terminal Evaluation Reports Rated as High Quality

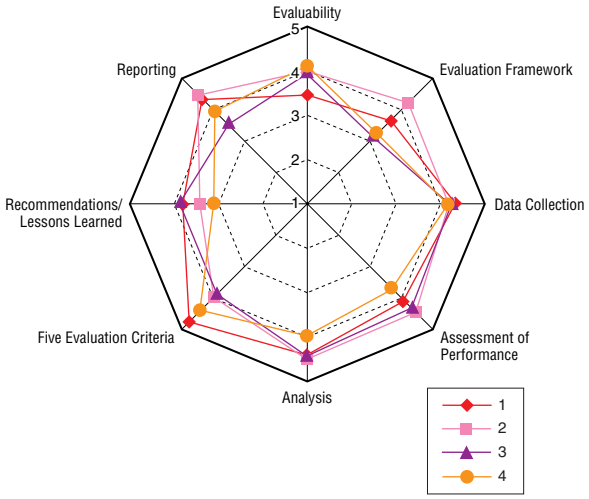
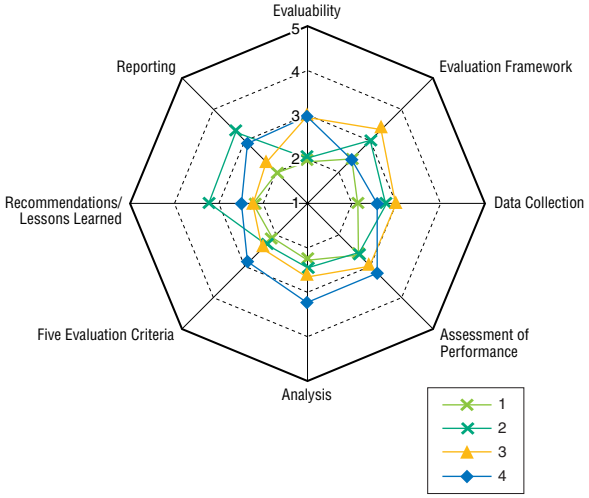


Figure 4-6 Worst Four Terminal Evaluation Reports Rated as Poor Quality



for evaluating the quality of reports.

The four evaluation reports rated as high quality are listed in Table 4-4, and those rated as poor quality in Table 4-5. The converted scores (out of five full scores) in the tables are the figures obtained by converting the total scores into five-scale ratings that correspond to the scores of each evaluation criterion.

The project whose terminal evaluation was rated the highest quality is the Maternal and Child Health Project in Cambodia (Phase 2). In Cambodia, not only medical facilities but also human resources were in a shambles due to a long-lasting civil war. To alleviate the situation, a project-type technical cooperation in the area of maternal and child health, which received little assistance from other aid agencies, was implemented from 1995 to 2000. After the termination of the project, the needs for improvement of maternal and child health services in rural areas was high, and the technical cooperation project was carried out with the purpose of human resource development for improving

Table 4-3 Average Scores of the Top Four and Worst Four Projects in the Total Scores

Evaluation Criteria	Average Scores		Difference in Average Scores
	Top four projects	Worst four projects	
Evaluability	3.89	2.53	1.36*
Evaluation Framework	3.59	2.77	0.81*
Data Collection	4.23	2.71	1.52**
Assessment of Performance	4.09	2.77	1.32**
Analysis	4.30	2.69	1.62**
Five Evaluation Criteria	4.22	2.51	1.71**
Recommendations/Lessons Learned	3.47	2.59	0.89*
Reporting	4.02	2.63	1.40**

* The difference in significance level between the top and worst four projects is 5% on average.
** The difference in significance level between the top and worst four projects is 1% on average.

maternal and child health. The project improved the quality of the National Maternal and Child Health Center and regional hospitals and developed leading medical professionals with appropriate knowledge and skills through training for doctors and midwives to establish a regional health system. As a result, support services were made available for other hospitals and the training program was accredited as a national training course by the Ministry of Health.

As the Ministry of Health instructed to implement a system for charging medical fees, which was first launched by this project, this system has spread throughout the country, generating institutional impact of the project.

The quality of the terminal evaluation of this project was determined to be high for the following reasons. Despite the drawback of indicators not being quantified, the framework of the project was clear and the alternative indicators were available to enable assessment. Data were collected from various sources and aspects, appropriate analyses were conducted from various perspectives and the reasoning was easy to follow.

The project whose terminal evaluation was rated the poorest in quality is the Project of Haraz Agricultural Human Resources Development Center in Iran. It is a technical cooperation project that worked to nurture engineers and provide education for farmers for the purpose of disseminating modern rice farming technology in the rice-producing areas along the Caspian coast. This project renovated and used the facility developed by the Haraz River Basin Agricultural Development Project, which ended in fiscal 1996, as the Haraz Agricultural Human Resources Development Center. It aimed to reinforce and improve the engineer training function of the Haraz Agricultural Human Resources Development Center. Fifty-two out of 59 courses scheduled kicked off, and 569 engineers, skilled workers, and leading farmers were enrolled as trainees (1,400 people capacity). However, at the time of the evaluation, the training center was not finished, training was not in full swing, and the development of teaching

materials was incomplete.

The following are the reasons why the quality of the terminal evaluation was poor. First, there was a defect in the reasoning of PDM represented by irrelevant relationships between project purpose and outputs. Without an evaluation grid, no data from beneficiaries were collected and no representative of the counterpart was interviewed; consequently data collection for analysis was insufficient. As for analysis, promoting and impeding factors were not fully analyzed. With regard to relevance, the consistency with Japan's aid policies was not touched upon. Effectiveness and efficiency were assessed high, which appeared unlikely from the analysis results due to the fact the training center and teaching materials had not been completed.

(4) Year-to-Year Changes in the Quality of Evaluation

We have thus far examined the quality of evaluation targeting 45 terminal evaluations in fiscal 2003 and fiscal 2004. In the last fiscal year, secondary evaluation of terminal evaluations were carried out for 38 projects in fiscal 2002 and 10 projects in fiscal 2003. Now, we will take a look at how the quality of terminal evaluation has changed over the years.

This year's secondary evaluators are different from those of last year. The evaluation criteria are the same but the viewpoints are somewhat different. From this, it is assumed that evaluation standards might be different, and it would not be possible to conduct year-to-year comparisons as they are. In addition, the rating system of evaluation criteria was based on a scale of 10 last year, but 5 this year. Thus, using the evaluation results of projects that were evaluated in both years, the evaluation scores of last year were converted to the standards of this year so that the evaluation standards would be the same, and then, a comparison analysis was performed.

In this way, the average scores of 38 projects in 2002 and 38 in 2003 and 17 in 2004 were obtained by evaluation criterion, and they are shown in Figure 4-7 and Table 4-6. Looking at any changes in the scores between fiscal 2002 and fiscal 2003 to see if the average scores have improved year to year, it is found that the difference in the average scores is quite small or even negligible. When comparisons are made between fiscal 2002 and fiscal 2004, the average scores of 2004 increased significantly for "evaluatability," "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting" from the statistical point of view. When comparisons are made between fiscal 2003 and fiscal 2004, the average scores of 2004 increased significantly for "data collection," "assessment of performance," "analysis," and "five evaluation criteria."

From above, it is fair to conclude that in fiscal 2004 the quality of terminal evaluation improved in all the evaluation criteria except "evaluation framework" and "recommendations/lessons learned."

Next, we will look at changes in distribution of the scores in each evaluation criterion. Figure 4-8 shows the percentage of reports with scores between 1.0 and 1.9, between 2.0 and 2.9, between 3.0 and 3.9, and between 4.0 and 5.0 in each evaluation criterion. From the figure, in fiscal 2004 the percentage of the reports with scores between 1.0 and 1.9 and between 2.0 and 2.9 decreased, while the percentage of reports with "average" rating and those with scores between 4.0 and 5.0 increased, indicating an improvement in the quality of terminal evaluation for fiscal 2004 compared with that for 2002 and 2003. In the criterion of "evaluation framework," there was no significant difference in average scores between fiscal years. The percentage of the scores between 3.0 and 3.9, and between 4.0 and 5.0 increased in fiscal 2004 in comparison with 2002 and 2003, while that of scores between 2.0 and 2.9 decreased, suggesting an improvement in the quality of terminal evaluation of 2004.

Table 4-4 Good Quality Terminal Evaluations

Country	Project Title	Total Score	Converted Score (out of five points)	Fiscal Year
1 Cambodia	Maternal and Child Health Project (Phase 2)	32.8	4.11	2004
2 Thailand	Reforestation and Extension Project in the Northeast of Thailand (Phase 2)	32.6	4.08	2004
3 Argentina	Regional Geological Mapping with Advanced Satellite Data	31.1	3.88	2004
4 Ethiopia	Laboratory Support for Polio Eradication: LAST Polio Project	30.6	3.83	2003

Table 4-5 Poor Quality Terminal Evaluations

Country	Project Title	Total Score	Converted Score (out of five points)	Fiscal Year
1 Iran	Project of Haraz Agricultural Human Resources Development Center	17.9	2.24	2003
2 Zambia	Technical and Vocational Training Improvement Project (Aftercare)	22.0	2.75	2003
3 China	Enhancement of Agricultural Extension System Project	22.0	2.75	2003
4 Cambodia	Improvement of the Survey and Forecast System on Meteorology and Agro-meteorolog	22.8	2.85	2003

Figure 4-7 Year-to-Year Changes in the Quality of Evaluations (Average Scores)

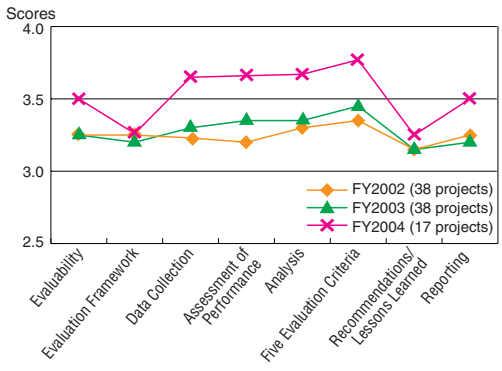
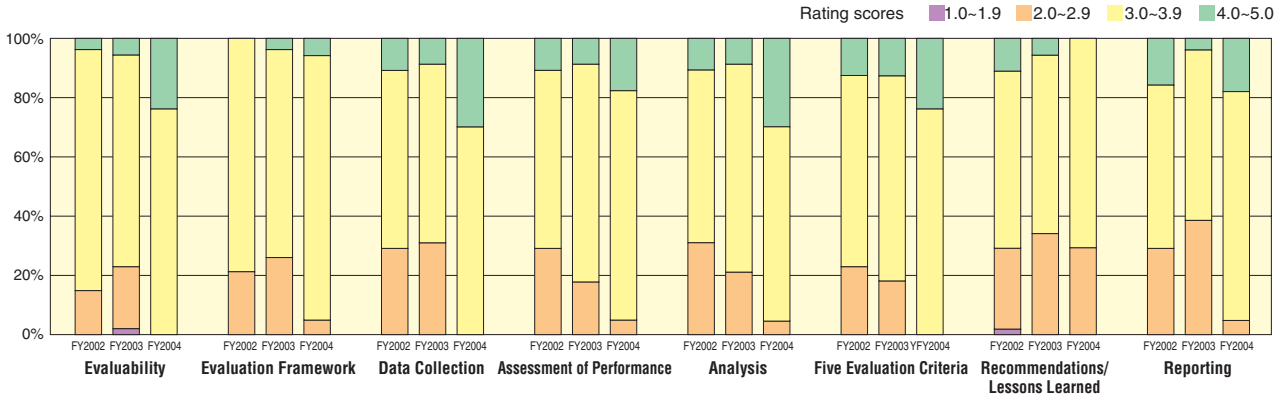


Table 4-6 Year-to-Year Changes in the Quality of Evaluations

Evaluation criteria	Average scores			Difference in the average between years		
	FY2002(A)	FY2003(B)	FY2004(C)	(B)-(A)	(C)-(A)	(C)-(B)
I Preconditions for Conducting Appropriate Evaluation						
Evaluability	3.26	3.17	3.51	– 0.09	0.25*	0.34
II Key Evaluation Criteria						
Evaluation Framework	3.26	3.14	3.29	– 0.13	0.03	0.16
Data Collection	3.26	3.25	3.64	0.00	0.38**	0.39**
Assessment of Performance	3.22	3.34	3.66	0.12	0.44*	0.32**
Analysis	3.30	3.28	3.67	– 0.02	0.38*	0.39*
Five Evaluation Criteria	3.37	3.36	3.78	0.00	0.41**	0.41**
Recommendations/Lessons Learned	3.16	3.15	3.27	0.00	0.12	0.12
Reporting	3.29	3.12	3.51	– 0.16	0.22*	0.39

* The difference in significance level between the scores in fiscal years is 5% on average.
** The difference in significance level between the scores in fiscal years is 1% on average.

Figure 4-8 Transition of the Quality of Terminal Evaluations over the Years (Distribution of Scores)



(5) Improving Primary Evaluation

■ Role of Terminal Evaluation

The main objectives of terminal evaluation are to accurately measure the degree of achievement of the goals and to obtain recommendations and lessons learned by considering future support systems through the verification of an implementation system in the partner country. However, some reports just listed justifications or excuses for terminating or continuing cooperation as the conclusions of terminal evaluation.

In some cases, termination of a project is decided because it generated remarkable outcome based on the reasonable result of terminal evaluation. In other cases, termination of a project is decided because it did not generate expected outcome, and the cooperation approach is to be reexamined. There may be a case where it takes time for outcomes to be evident even though the cooperation approach is appropriate. In such a case, it is not wise to terminate the cooperation. For that reason, one may justifiably conclude that the project should continue.

In any case, such decisions should be made after evaluation results are released. The writing of the report itself should be limited to the evaluation results of the evaluators and the presentation of recommendations based on the results.

■ Viewpoint for Assessment of Impact

Since terminal evaluation is conducted several months prior to the actual termination of a project, it seems to be a viable approach to evaluate the degree of achievement of the purpose, relevance, and efficiency, and to compile recommendations about sustainability through the assessment of the implementation system of the partner country. However, in most cases, the impact of the project would be limited to the prediction as to how much impact (positive/negative) would be brought about in the future. Even in that case, it is necessary to discover an impact, however small it may be, to present the basis for the prediction and to increase the credibility, instead of presenting wishful thinking.

■ Timing of Terminal Evaluation

The last six months or so of a project is the time to finalize various activities. One may argue that a pursuit of the direction of finishing is a hidden objective of terminal evaluation. It does not mean that it is inappropriate; however, if the objective is different, the direction of evaluation is different, too. Therefore, it is essential to clarify the objective of terminal evaluation when determining the timing of terminal evaluation.

■ Survey on Beneficiaries

When collecting data, some survey targets (questionnaire surveys, interviews, etc) were extremely limited to a small group of people, such as counterpart members of the implementing organizations and trainees. In order to verify relevance and the

project implementation process, it is important to extend the data sources from the policy-making level to beneficiaries and the people in the vicinity.

■ Understanding Important External Assumptions

It is necessary to improve the understanding of items listed as important external assumptions. When analyzing the effectiveness of a project, confirmation of the degree of achievement of the purpose alone is not enough. In order to determine how much the project itself has contributed to the achievement of the purpose, it is inevitable to fully understand internal and external factors that may greatly influence the outcome of the project.

■ How to Write a Report

It is necessary to present guidelines for writing a report. The recommended style is that a report should start with a summary that can be understood by the general public, followed by the main text, and raw data should be attached at the end. Some efforts to simplify the report were observed; however, in some reports, the same sentences were repeated in the executive summary, the text, and conclusion. In many reports, the statement in the text simply says that the quality of trainees improved, giving no basis for that statement, and one needs to look into the evaluation grid to find the details (in some cases, the details are still unclear from the evaluation grid). It is better to present more detailed description in the text, such as data indicating the basis for important items. In fact, in some cases, a field survey report from consultants presented as an appendix explained more clearly about the conditions and was more interesting. The basis for a judgment should be furnished and data should be attached at the end.

Some reports contained a large number of supplementary documents while the text itself was short. Many documents were not referred to in the text, and therefore thought to be unnecessary. It is not always a waste to include such documents, because, otherwise, they could be dispersed and lost. However, they have little to do with the evaluation or analysis, and therefore, it would be better to exclude a line-up of facts and lists (the counterpart's schedule of a trip to Japan, list of equipment, etc.) from the report, and include them only in a CD-ROM.

What is most important about the report is that it be understandable to non-experts. Particularly in an area like basic research and medicine, when an achievement is made in an experiment, it would be understandable for non-experts if an explanation is given as to how much impact the achievement has from a broader perspective; for example, success in the cultivation of a microorganism would be the first step to the development of a drug for early detection of a disease. It will be more understandable if the report states how long it would take to actually develop the drug from the cultivated microorganism.

(6) Summary of the Quality of Primary Evaluation

The quality of terminal evaluation satisfies a certain level; however, the scores for "evaluation framework" and "recommendations/lessons learned" are lower than those for other criteria. First, the participation of partner countries in evaluation is rated low. This seems to be because the participation of the partner country at the time of evaluation was vague, and the report does not clarify this point. In order to improve the quality of evaluation, it is necessary to increase the participation of partner countries and stipulate the participants' relationships with the project and evaluation method in the report to ensure neutrality and specialty of evaluation participants.

It was found that "recommendations/lessons learned" were not fully extracted from the results of analysis. Some of the statements of the team leader contain useful recommendations/lessons learned. Since recommendations/lessons learned are useful for improving the effect of projects when implementing similar projects in the future, it is necessary to devise a way to deal with the recommendations/lessons learned that have not been agreed upon with the partner countries.

According to the analyses thus far, the quality of evaluation results have improved over the years. In particular, the difference is large between the projects evaluated in fiscal 2003 and the projects in fiscal 2004. One of the reasons may be that the JICA Guidelines for Project Evaluation were drastically revised in February 2004 based on the secondary evaluation of fiscal 2003 and the efforts to improve the quality of primary evaluation were made in line with the new guidelines. Due to progress in field based management, the evaluation system is also undergoing institutional change, in which more evaluations are conducted at overseas offices. The fact that the new evaluation guidelines have contributed to the improvement of the quality of evaluation indicates the feedback of the secondary evaluation results is effective for improving the quality of primary evaluation. Moreover, ex-ante evaluation was introduced in fiscal 2001, and among projects which went through ex-ante evaluation the terminal evaluations for three-year projects were conducted in fiscal 2004. Thus, the introduction of ex-ante evaluation may have had some impact on the quality improvement. Nonetheless, there are only two cases of such evaluation in this year's analysis, and the effects of ex-ante evaluation on the improvement of quality of terminal evaluation need to be further investigated next year.

Furthermore, JICA, in order to improve the quality of terminal evaluation, established the JICA Good Practice Evaluation Award in fiscal 2004 to recognize those evaluations that serve as a model for other projects, and this award uses the secondary evaluation results. It is of particular significance now that it is apparent that the secondary evaluation results can contribute to the quality of evaluation by combining the advantages of both external and internal evaluations.

1-3 Project Evaluation by Secondary Evaluators Based on Terminal Evaluation Reports

(1) Summary of the Secondary Evaluation of Projects

We conducted secondary evaluation on 28 projects evaluated in fiscal 2003 and 17 projects evaluated in fiscal 2004 using the terminal evaluation reports from the perspective of the DAC Evaluation Criteria. Figure 4-9 shows the result of the project evaluation gleaned from the reports by the secondary evaluators.

All the average scores for the projects are in the 3-point range, being above the level of "average." Among the five evaluation criteria, the average scores for "relevance" and "effectiveness" are high with 3.6 points and 3.5 points, respectively, while the average scores for "efficiency," "impact," and "sustainability" are relatively low with around 3.2 points.

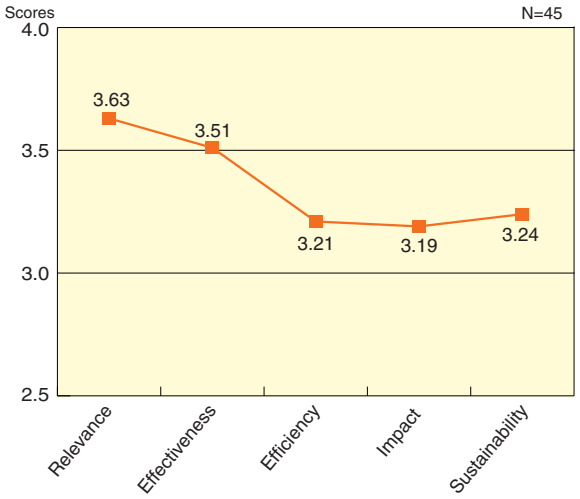
Figure 4-10 shows the distribution of scores for the projects' evaluation. According to the distribution, scores are clustered around the range of 3.0 to 3.9 points for every criterion. Most of the scores for "relevance" are 3.5 points or higher and none are below 2.5 points. The scores for "efficiency" and "impact" are clustered in the range of 2.5 to 3.9, and the rating is rather low. The variances of the scores for "effectiveness" and "sustainability" are large. Most scores for "effectiveness" fall in the range between 3.0 and 3.9, with a few points above 4.5. Most scores for "sustainability" are below 3.0; some are above 4.0, while some are below 2.0, showing low ratings.

Additionally, we conducted secondary evaluation from various viewpoints for each evaluation criterion based on the information obtained from the reports. Figure 4-11 shows the average scores for the viewpoints. The scores for all the viewpoints under the criterion of "relevance" are generally high. Among them, the average score for "validity" is the highest with 3.8, indicating a high relevance with respect to project implementation. "Appropriate approach," which is concerned with project design, was rated the lowest, but the average score of 3.4 is still high compared with viewpoints in other criteria.

In the criterion of "effectiveness," "achievement level of project purpose" received a higher average score than "causal relationships between outputs and project purpose." In the criterion of "efficiency", the average score for "appropriate implementation process" is higher than that for "cost-effectiveness" which determines if efforts were made to achieve more outcomes with lower costs.

In the criterion of "impact", the average scores for "achievement level of overall goal," "causal relationships between project purpose and overall goal," and "unintended positive and negative impact" are more or less the same. Among the five viewpoints of "sustainability," "organizational sustainability," which is associated with the organizational strength to ensure sustainability, scored the highest points, whereas the average score for "financial sustainability," which is related to financial capacity to ensure sustainability, scored the lowest. All the average scores for the viewpoints are above 3.0 points except for "financial sustainability," which scored below 3.0.

Figure 4-9 Project Evaluation by Secondary Evaluators



(2) Evaluation of Projects by Fiscal Year, Region, and Sector

1) Evaluation by Fiscal Year

Figure 4-12 shows the year-to-year change in average scores; in other words, how the project evaluation has changed over the years. Table 4-7 shows the result of statistical analysis to examine if the average scores are different by fiscal year. As clearly shown in Figure 4-12 and Table 4-7, the average scores for fiscal 2004 in every evaluation criterion are higher than those for fiscal 2002 and fiscal 2003, suggesting that projects are improving, though the judgment was based on limited information available in the reports.

2) Evaluation by Sector

Next, we will take a look at project evaluation by sector. The evaluation targeted 15 projects in the sector of health and medical care, 10 in social development, 10 in agricultural development, three in mining and industrial development, and seven in forest and natural environment. The projects in health and medical care include medical education, strengthening of regional health, and improvement of medical technology at hospitals. The projects in social development include regional development, vocational training and improvement of marine education. The projects in agricultural development include irrigation technology, agricultural technology development, and productivity improvement. The projects in mining and industrial development include electricity technology, casting technology, and industrial water technology development. The projects in forestry and natural

Figure 4-10 Distribution of Evaluation Scores for Projects by Secondary Evaluators

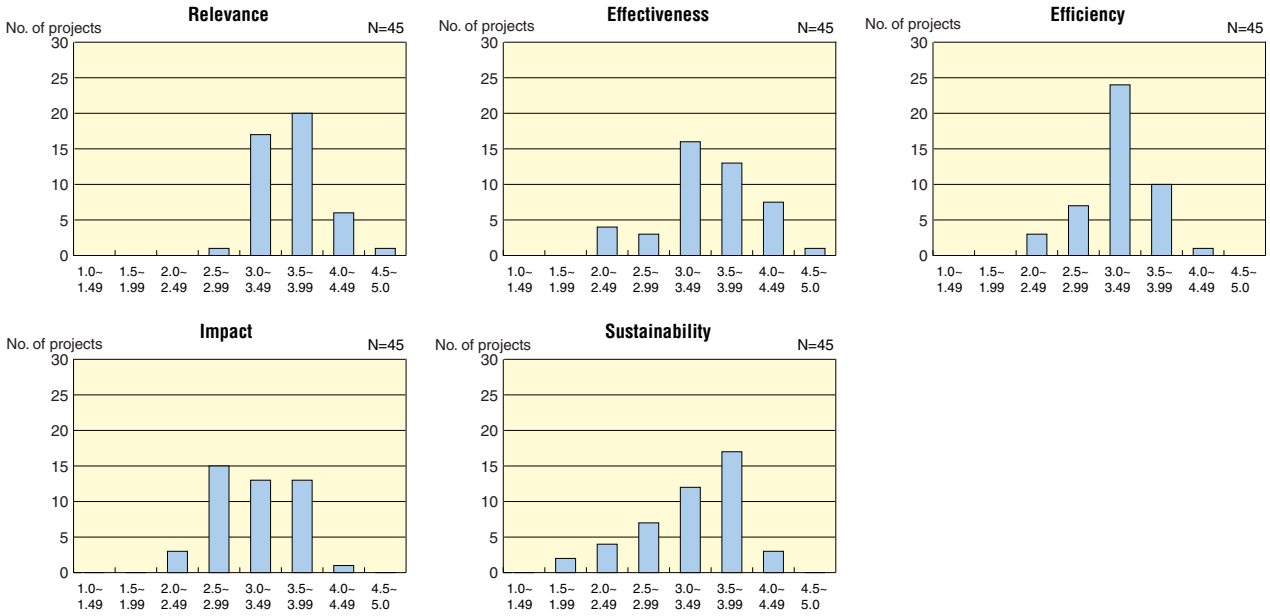


Figure 4-11 Project Evaluation by Secondary Evaluators: Viewpoints (Average Scores)

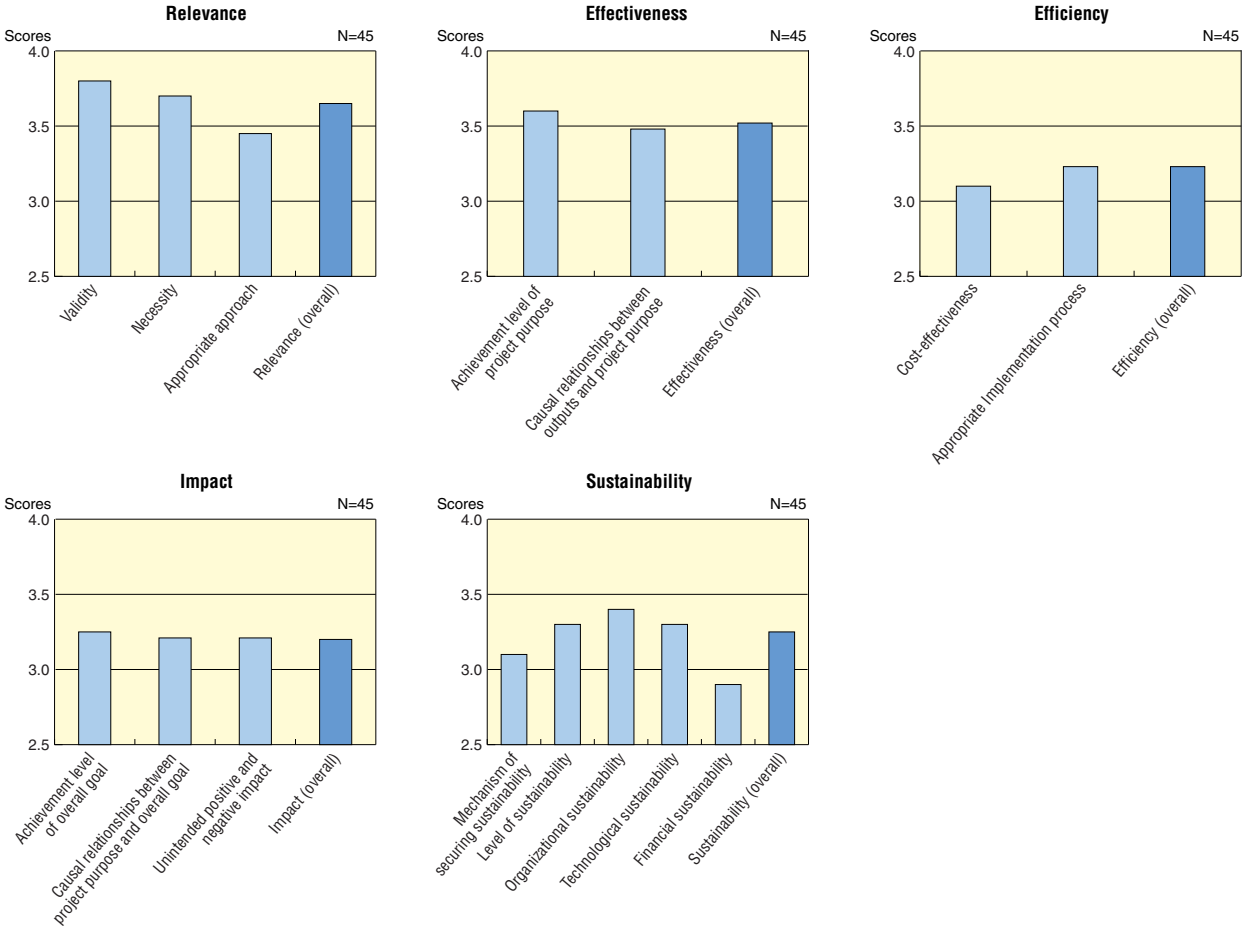


Figure 4-12 Year-to-Year Changes for Project Evaluation by Secondary Evaluators (Average Score)

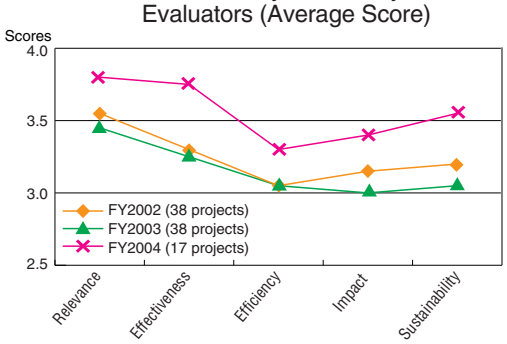


Table 4-7 Year-to-Year Changes for Project Evaluation by Secondary Evaluators

	Average Score			Difference in Average Scores		
Evaluation Criteria	FY2002(A)	FY2003(B)	FY2004(C)	(B)-(A)	(C)-(A)	(C)-(B)
Relevance	3.53	3.46	3.85	– 0.08	0.31**	0.39*
Effectiveness	3.34	3.28	3.75	– 0.06	0.41**	0.47*
Efficiency	3.07	3.08	3.36	0.00	0.28*	0.28*
Impact	3.11	3.00	3.44	– 0.11	0.33**	0.44*
Sustainability	3.17	3.05	3.53	– 0.12	0.35**	0.47*

* The difference in significance level between the scores in fiscal years is 5% on average.
** The difference in significance level between the scores of fiscal years is 1% on average.

environment include aquaculture promotion, environmental conservation, forestry study, and forestation.

Figure 4-13 shows the project evaluation by sector. Though the number of projects varies by sector, the evaluation results exhibit differences by sector. Although there is no statistically significant difference, the average scores for the mining and industrial development sector tend to be generally higher than that for other sectors in every criterion and the average of the agricultural development sector tend to be lower. With respect to "impact," there is a statistically significant difference between the average scores for the mining and industrial sector and the agricultural development sector, and the mining and industrial sector was rated high because more impacts had been identified in the mining and industrial sector than the agricultural development sector. In the agricultural development sector, the average score for "sustainability" is below "average" with 2.9 points.

The reasons why the ratings for the agricultural development sector are low and those for the mining and industrial sector are high may be attributed to the nature of cooperation. In specific terms, many projects in the agricultural development sector are provided over the long-term, are hard to focus on, and take a long time to achieve the goal, whereas those in the mining and industrial sector can achieve objectives in a relatively short period of time.

All three projects in the mining and industrial sector received the average score of 3.0 points or higher. Regional Geologic Mapping with Advanced Satellite Sensors in Argentina was a technical cooperation project aimed at facilitating geological mapping by advanced satellite data processing and analysis technology and introducing necessary equipment and software. Using such data as ASTER or PALSAR, precise geological and thematic mappings were constructed, and furthermore, due to increased interests in the use of ASTER data in the areas such as mine resource development and oil development, orders for its graphic images came in. The average scores for this project received high ratings: "relevance," "effectiveness," and "impact" are 4.0 points or higher, "efficiency" and "sustainability" are 3.7 points or higher.

The evaluation on projects in the agricultural development sector varies. The Joint Study on Biological Control of Soil-borne Plant Diseases in Argentina aimed to develop biological control method against soil-borne plant infectious diseases and as a result, developed two types of very effective comprehensive control methods. Improvement of Productivity for the Small-scale Dairy Farmers Project in the Republic of Chile established a training center to improve the productivity of small-scale dairy farms and support engineers and farmers to acquire and utilize knowledge and skills for dairy farming production. The average scores for the evaluation criteria of these two projects are 3.3 points or higher, many of which are 3.6 or higher and those for "relevance" and "effectiveness" are above 4 points. On the other hand, the Project of Haraz Agricultural Human Resources Development Center in Iran received the scores for all the criteria between 2.0 and 2.9 on average. Promotion of Sustainable Community Based Small-holder Irrigation in Kenya was carried out with the purpose of stabilizing agricultural production through the development of small-scale irrigation initiated by the farm community. This project, which was low in efficiency and feasibility due to insufficient input, was given 3.1 points for "relevance", but the scores for "effectiveness" and "impact" are between 2.0 and 2.9 and those for "efficiency" and "sustainability" are between 1.0 and 1.9. The average scores for other projects with low ratings were in the 2-point range for "effectiveness," "efficiency," "impact," and "sustainability."

3) Evaluation by Region

The number of projects subject to the secondary evaluation is 25 in Asia and Oceania, which is the largest, followed by nine in Latin America, eight in Africa, and three in Middle East. Figure 4-14 shows the outcomes of projects by region. The number of projects in regions other than Asia and Oceania is small, and no projects evaluated in 2004 are included in Africa; there are some differences in the situations by region.

Statistically significant differences were observed between regions. The average scores of Middle East for "relevance," "effectiveness," and "efficiency" are lower than those of Latin America and Asia and Oceania. The average scores for "efficiency" in Middle East and Africa are between 2.0 and 2.9, which

Figure 4-13 Evaluation by Sector (Average)

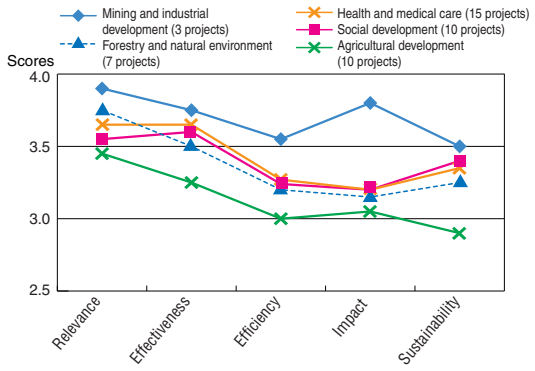
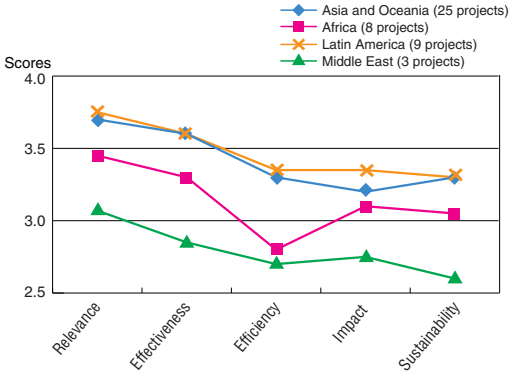


Figure 4-14 Evaluation by Region (Average)



are lower than those in Latin America and Asia and Oceania. In general, the average scores in Middle East tend to be lower than in other regions, with the scores for four criteria except "relevance" between 2.0 and 2.9. The projects in Middle East are: the Project on Improvement of Maritime Education in Turkey, the Tuberculosis Control Project (Phase 3) in Yemen, and the Project of Haraz Agricultural Human Resources Development Center in Iran. The average scores of the Project of Haraz Agricultural Human Resources Development Center in Iran are between 2.0 and 2.9 in all evaluation criteria, and those for "effectiveness" and "efficiency" are relatively low with 2.3 points or lower. In the Project on Improvement of Maritime Education in Turkey, technical cooperation was provided to establish a maritime education system that satisfies international standards and to train sailors. As a result, the organizational capacity for the operation of educational equipment and designing and implementation of training content improved to some extent, but some equipment was not used effectively. The average scores for "relevance" and "effectiveness" of this project are 3.0 points, "efficiency" and "impact" are 2.7 or higher; however, the average score for "sustainability" is low with 2.3 points. The Tuberculosis Control Project (Phase 3) in Yemen carried out the national tuberculosis control programs throughout the country and the area covered by this project expanded to 98% of the country; however, one step was away from achieving a success rate of treatment. The average scores for "relevance," "effectiveness," "efficiency," and "impact" in this project are between 3.0 and 3.9, but the average score for "sustainability" is 2.9 points.

(3) Overall Evaluation by Secondary Evaluators

Based on the aggregates of scores for the five evaluation criteria, which the secondary evaluators gave by judging from the reports (5 to 25 points), the projects were classified into four categories: excellent (20 points or higher), good (15-19 points), poor (10-14 points) and very poor (5-9 points). Figure 4-15 shows the percentage of each category in every fiscal year.

In fiscal 2002, there are some projects in the category of "very poor," but there are none in that category in fiscal 2003 and fiscal 2004. In fiscal 2004, fewer projects are "poor" and the percentage of "excellent" increased. It is suggested that many projects in fiscal 2004 are successful, although not definite, since the projects subject to the secondary evaluation in fiscal 2004 are just a part of all the projects implemented in fiscal 2004.

Among 45 projects subject to the evaluation of fiscal 2003 and 2004, four projects with high scores of 20 points or higher for five evaluation criteria and four projects with low scores of 13 points or lower were selected and are shown in Figure 4-16 and Figure 4-17. All the four projects with high scores are those in fiscal 2004. The lowest total score of the projects in fiscal 2004 is 14 points (see Table 4-10). All the four projects with low scores are those in fiscal 2003. Table 4-8 shows the difference in the averages scores between the four projects with high total scores and the four with low total scores. There is statistically significant difference between the two groups in each evaluation criterion.

As clear from Figures 4-16 and 4-17 and Table 4-8, among five evaluation criteria, the projects with high total scores received high points for "relevance," "effectiveness," and "sustainability" with the average score of 4.1 or higher; in particular, "effectiveness" received the highest score. Although "efficiency" is less than 4.0, there is little variance among the projects. The projects with low total scores received relatively high points for "relevance" and "impact." However, the difference between the two groups shows that there is a large difference in the average scores for "effectiveness," "efficiency," and "sustainability." From these results, it is believed that efficiently implemented projects with a

Figure 4-15 Overall Project Evaluation by the Secondary Evaluators

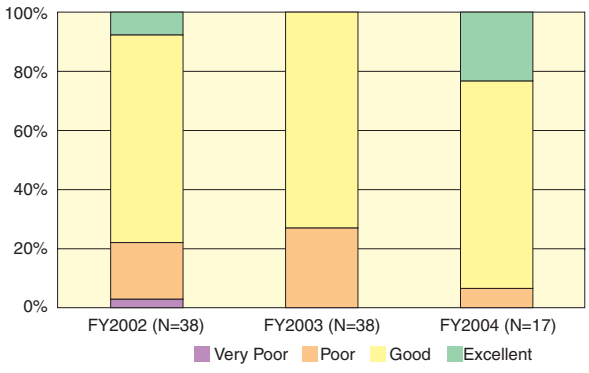


Figure 4-16 Top Four Projects

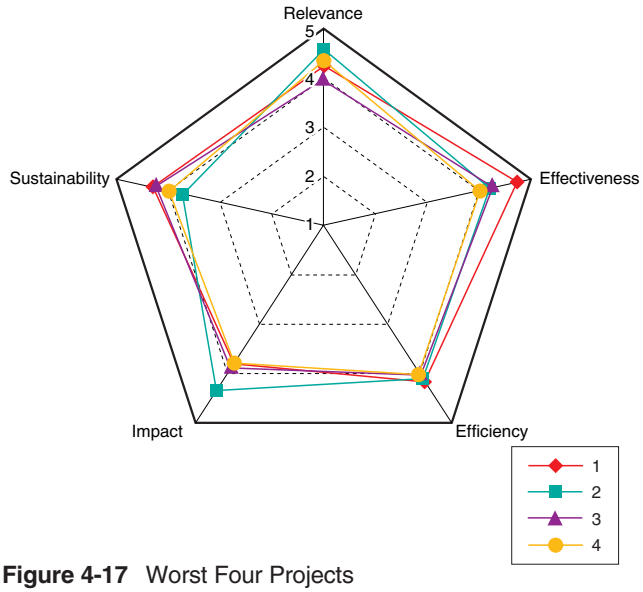
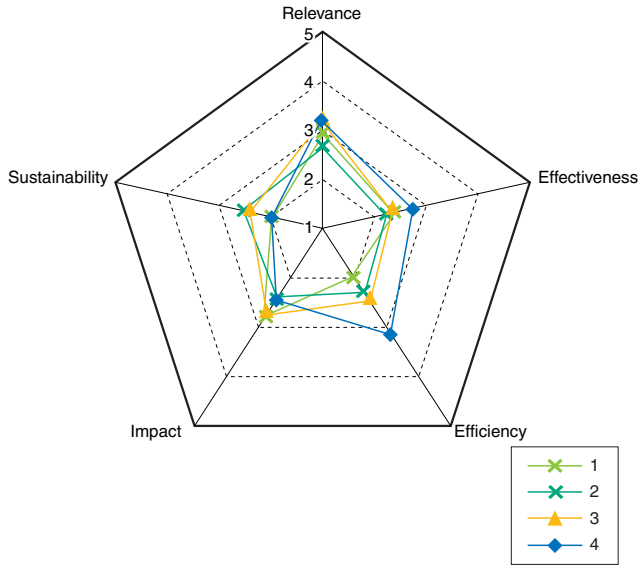


Figure 4-17 Worst Four Projects



high level of goal achievement and high sustainability promises high evaluation.

So what happened between fiscal 2003 and fiscal 2004? The PCM method, for instance, was introduced in fiscal 1996, but the position of PDM was not clearly defined and no flexible modification was made at the implementation stage until fiscal 2002. Through the revision and dissemination of the Evaluation Guidelines, the PDM has been modified since fiscal 2003, along with the changes in project planning.

With respect to PDM, many of the projects that terminated by fiscal 2002 formulated a revised version of PDM, or PDMe, at the final stage of the project and conducted evaluation. Since the revised Evaluation Guidelines clearly refer to flexibility of the PDM, the stakeholders renewed their awareness and the number of PDM, which is revised accordingly during the project, is on the rise. This would increase evaluability and eventually lead to the improvement of the assessment of results. A full fledged assessment on this issue will be conducted from next year on.

Table 4-8 Average Scores of Top and Worst Four Projects

Evaluation Criteria	Average Scores		Difference in Average Scores
	Top four projects	Worst four projects	
Relevance	4.26	3.08	1.18*
Effectiveness	4.41	2.43	1.98*
Efficiency	3.95	2.42	1.53*
Impact	3.89	2.59	1.30*
Sustainability	4.11	2.24	1.87*

* The difference in significance level between the average scores of the top and worst four projects is 1%.

As Table 4-9 shows, four projects were rated "excellent" by the secondary evaluators. No projects were rated "very poor" and the projects that received "poor" are shown in Table 4-10. The converted scores (out of five full scores) in the tables are the figures obtained by converting the total scores into five-scale rating to correspond to the scores of each evaluation criterion.

The project that received the highest score is the Research Project on Timber from Man-made Forests in China. This project was technical cooperation for research aiming at sustainable and effective utilization of man-made wood in order to harmonize the demand for timber associated with social development in China and environmental issues. The purpose of the project was to strengthen the capacity of basic research on man-made wood at the Research Institute of Wood Industry, Chinese Academy of Forestry. Counterparts of the project were in high ranking positions, and had college degrees or were enrolled in post-graduate courses. In addition, students have received or are enrolled to receive degrees in man-made wood production. The consolidated technology transfer increased the organizational power, leading to the high rating in efficiency. The project purpose was achieved as expected, proving high effectiveness. These factors led to the expansion of the research organization and new budgetary measures, thus generating high sustainability. These are the basis for the high score.

The project called Promotion of Sustainable Community Based Small-holder Irrigation in Kenya was given the lowest total score. The purpose of the project was to develop small-scale irrigation managed by farmers to stabilize agricultural production. There was a lack of common awareness about project design and purpose and insufficient input, such as dispatch of experts and operation costs, leading to low efficiency. Sufficient outcomes were not generated and financial sustainability was extremely low. These are the basis for low scores.

(4) Relationships between the Results of Project Evaluation and the Results of Evaluation of the Quality of Terminal Evaluation

Figure 4-18 illustrates the relation between the total scores of project evaluation carried out this year on a total of 45 projects in

2003 and 2004 and the total scores of terminal evaluations in quality. The relation between the project evaluation and the quality of terminal evaluation shows that two out of four projects that were rated "excellent" by the secondary evaluators were also rated high in terms of quality of terminal evaluation as described in section 1-2-(3): namely, Regional Geologic Mapping with Advanced Satellite Sensors in Argentina (A in the figure) and the Maternal and Child Health Project in Cambodia (Phase 2) (B in the figure). In addition, three projects that were rated "poor" by the secondary evaluators were also rated low in terms of quality of terminal evaluation (Table 4-5): namely, Improvement of the Survey and Forecast System on Meteorology and Agro-meteorology in Cambodia (C in the figure), Technical and Vocational Training Improvement Project (Aftercare) in Zambia (D in the figure) and the Project of Haraz Agricultural Human Resources Development Center in Iran (E in the figure).

Table 4-11 shows the correlation between the scores for criteria of project evaluation and the scores for criteria of the quality of terminal evaluation. "Relevance" of projects is closely correlated with "evaluability," "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting." "Effectiveness" of projects is also closely correlated with "evaluability," "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting," and its correlation with "evaluatbility" and "five evaluation criteria" is particularly close. "Efficiency" is closely correlated with "evaluability," "analysis," and "five evaluation criteria." "Impact" is closely correlated with "evaluability," "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting." "Sustainability" is closely correlated with "five evaluation criteria." Every criterion of project evaluation is closely correlated with "evaluability." In conclusion, there is close correlation between project evaluation and the quality of terminal evaluation.

A good project has clearly purpose and goal, carries out various activities as planned, and properly manages data of monitoring each time. These factors are believed to lead to effective project implementation, and facilitate assessment of performance and analysis to verify the effectiveness, which would then result in appropriate evaluation. This would also enable the preparation of clear and easy-to-understand evaluation reports.

The straight line (2) in Figure 4-18 indicates an expected value that estimates the quality of terminal evaluation based on project evaluation. The projects rated high tend to have high scores for the quality of terminal evaluation. However, as clearly shown in the figure, in some cases the quality of terminal evaluation deviates greatly from the average expected value estimated from the project evaluation. This actually refers to two types of projects. When the total scores for the quality of terminal evaluation exceed the expected value of the quality of terminal evaluation, the quality of terminal evaluation of the project is higher than what is estimated from project evaluation. And, when the total scores for the quality of terminal evaluation is lower than the expected value, the quality of terminal evaluation of the project is lower than what is estimated from project evaluation.

Based on the correlation between the project evaluation and the quality of terminal evaluation, the projects are classified into three groups, in order to probe the correlation between the project evaluation and the quality of terminal evaluation. When the total score for the quality of terminal evaluation is higher than the average expected value by 1.5 points, the project is classified as "a. the projects whose quality of terminal evaluation is higher than the expected value estimated from project evaluation," which represents the area above the line (1) in Figure 4-18. When the quality of terminal evaluation is within the range of ±1.5 of the expected value, the project is classified as "b. the projects whose project evaluation corresponds to the quality of terminal evalua-

Table 4-9 Projects Rated "Excellent" by the Secondary Evaluators

Country	Project Title	Total Score	Converted Score (out of five points)	Fiscal Year
1 China	Research Project on Timber from Man-made Forests	21.1	4.22	2004
2 Argentina	Regional Geologic Mapping with Advanced Satellite Sensors	20.9	4.18	2004
3 Viet Nam	The Project for Strengthening Training Capacity for Technical Workers in the Hanoi Industrial College	20.4	4.08	2004
4 Cambodia	The Maternal and Child Health Project (Phase 2)	20.2	4.03	2004

Table 4-10 Projects Rated "Poor" by the Secondary Evaluators

Country	Project Title	Total Score	Converted Score (out of five points)	Fiscal Year
1 Kenya	Promotion of Sustainable Community Based Small-holder Irrigation	12.2	2.45	2003
2 Iran	The Project of Haraz Agricultural Human Resources Development Center	12.3	2.47	2003
3 Zambia	Technical and Vocational Training Improvement Project (Aftercare)	13.2	2.64	2003
4 Cambodia	Improvement of the Survey and Forecast System on Meteorology and Agro-meteorology	13.3	2.66	2003
5 El Salvador	The Project on the Aquaculture Development	13.5	2.70	2003
6 Brazil	The Technological Development Project for Sustainable Agriculture in Eastern Amazonia	13.6	2.72	2003
7 Turkey	The Project on Improvement of Maritime Education	14.0	2.80	2004
8 Indonesia	Malaria Control in Lombok and Sumbawa Islands	14.2	2.85	2003
9 Nepal	Community Development and Forest/Watershed Conservation Project (Phase 2)	14.5	2.91	2003
10 Indonesia	The Mangrove Information Center Project	14.6	2.93	2003

Table 4-11 Correlation between Project Evaluation by the Secondary Evaluators and Evaluation of the Quality of Terminal Evaluation

Project Evaluation	Quality of Terminal Evaluation							
	Evaluability	Evaluation Framework	Data Collection	Assessment of Performance	Analysis	Five Evaluation Criteria	Recommendations /Lessons Learned	Reporting
Relevance	0.538**	0.340 *	0.523**	0.551**	0.501**	0.571**	0.263	0.455**
Effectiveness	0.712**	0.341 *	0.594**	0.542**	0.533**	0.647**	0.059	0.520**
Efficiency	0.574**	0.342 *	0.399**	0.388**	0.446**	0.550**	0.189	0.339 *
Impact	0.526**	0.331 *	0.520**	0.500**	0.483**	0.477**	0.107	0.460**
Sustainability	0.355 *	0.285	0.392**	0.302 *	0.391**	0.466**	0.176	0.389**

Figures show the coefficient of correlation *5% of significance level **1% of significance level N=45

Figure 4-18 Relationships between Project Evaluation by the Secondary Evaluators and Evaluation of the Quality of Terminal Evaluation

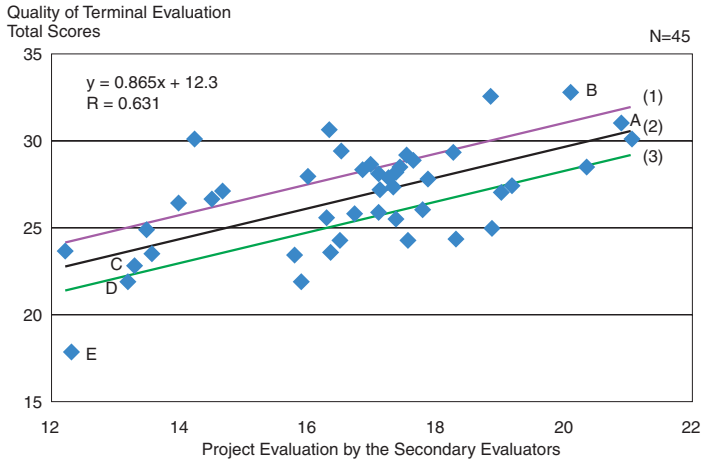


Figure 4-19 Evaluation of the Quality of Terminal Evaluation (By Group)

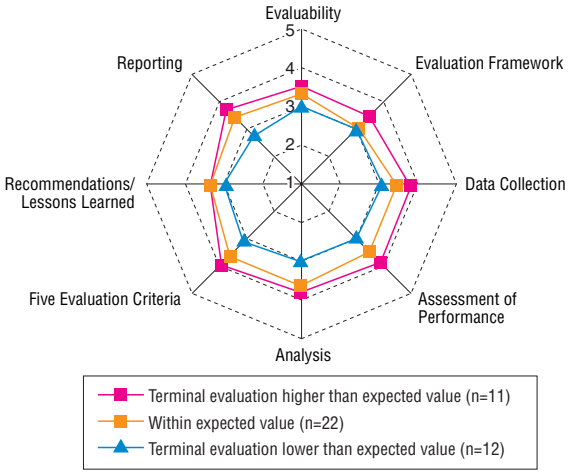


Table 4-12 Scores for Evaluation Criteria of the Reports (Average)

Evaluation Criteria	Average Scores			Difference in Average Scores		
	(A) Terminal evaluation higher than expected value (n=11)	(B) Within expected value (n=22)	(C) Terminal evaluation lower than expected value (n=12)	(A)-(B)	(C)-(B)	(A)-(C)
Evaluability	3.55	3.36	3.01	0.19	– 0.35*	0.54**
Evaluation Framework	3.41	3.13	3.04	0.28*	– 0.09	0.37**
Data Collection	3.78	3.46	3.00	0.32*	– 0.46**	0.78**
Assessment of Performance	3.78	3.51	3.02	0.26*	– 0.49**	0.76**
Analysis	3.73	3.56	2.98	0.17	– 0.58**	0.75**
Five Evaluation Criteria	3.93	3.52	3.14	0.41*	– 0.38*	0.79**
Recommendations/Lessons Learned	3.37	3.31	2.90	0.06	– 0.41**	0.47**
Reporting	3.67	3.39	2.76	0.28	– 0.64**	0.92**

* The difference in significance level between the average scores of groups is 5%.
** The difference in significance level between the average scores of groups is 1%.

tion," which represents the area between the lines (1) and (3). When the total score of the quality of terminal evaluation is lower than expected value by 1.5, the project is classified as "c. the projects whose quality of terminal evaluation is lower than the expected value estimated from project evaluation," which represents the area below the line (3) in the figure. Figure 4-19 shows the average scores of evaluation criteria of the three groups and Table 4-12 shows the average scores for evaluation criteria and differences in the average scores of the three groups.

As clearly shown in Table 4-12, the projects whose quality of terminal evaluation is lower than the expected value estimated from project evaluation generally have statistically significantly lower average scores for terminal evaluation than the average expected value; in particular, the difference is large in the average scores for "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting" ((C)-(B) in the table). The projects whose quality of terminal evaluation is higher than the expected value estimated from project evaluation generally have statistically higher average scores for "evaluation framework," "data collection," "assessment of performance," and "five evaluation criteria" ((A)-(B) in the table). Next, when the projects whose quality of terminal evaluation is higher than the expected value estimated from project evaluation are compared with the projects whose quality terminal evaluation is lower than the expected value estimated from project evaluation, there is a significant difference in the average scores for every evaluation criterion. In particular, the difference is large for "data collection," "assessment of performance," "analysis," "five evaluation criteria," and "reporting," and the projects with high quality terminal evaluation have high scores for these criteria.

Also as shown in the distribution of Figure 4-18, the projects with high evaluation scores are generally rated high in the quality of terminal evaluation, whereas the projects with low evaluation scores exhibit a large variance, with a large difference in the quality of terminal evaluation between the high scores and low scores.

From these results, good projects tend, in general, to have high quality of terminal evaluation. Nonetheless, regardless of the quality of projects, it is necessary when conducting evaluation projects to collect appropriate data from various aspects, assess and analyze the performance objectively and accurately, and clearly describe details using tables and figures. This enables the compilation of high quality reports.

The secondary evaluation result has revealed that not only the quality of terminal evaluation increased, but also the projects improved in fiscal 2004. Evaluation is a mechanism for quality assurance of projects and an increase in the quality of evaluation is expected to have a positive impact on the effective and efficient implementation of projects. In view of tasks necessary for evaluation to have a positive impact on the quality of projects, it is more relevant to consider that long-term efforts for strengthening evaluations, such as the introduction of the secondary evaluation, are the factors contributing to the positive impact on effectiveness of projects, instead of the introduction of the secondary evaluation resulting in immediate improvement of projects.

For example, such changes within JICA as enhancement of country- and issue-specific approaches, development of JICA Country Programs, introduction of ex-ante evaluation, and delegation of authorities to overseas offices seem to have influenced the improvement of the quality of projects. It is assumed that rather than influence by individual plans, what improves projects as a whole is sufficiently addressing issues and measures through the discussions based on the plans and implementing projects with evaluation in mind. Still, further analysis covering longer spans is warranted to verify specifics.

(5) Improving the Quality of Projects

■ Improving PDM

Monitoring the progress of projects and terminal evaluation are carried out based on PDM. From the experience with this year's secondary evaluation, some of the principles for good PDM were drawn out as follows.

- Target groups are clearly identified.
- Project purpose responds accurately to the needs of the local community.
- Indicators corresponding to the project purpose and the target values are clear.
- An agreement upon the PDM is reached among the parties concerned.
- All the important items are fully covered, accurately reflecting the reality of the project.
- Contents are understandable to the general public.
- Causal relationships in the process from the input to the overall goal is appropriately demonstrated.
- The responsible personnel for the PDM are always identifiable.

Although these points are nothing new, many of them have not been fully observed. This year's secondary evaluation saw that some projects were supply-driven, in which project purposes appeared to have been determined as a consequence of the limitation in the resources of the Japanese side and not based on the needs of the local community. There were also some projects whose contents of output and outcome (project purpose) did not show causal relationships. In another project, since the description of project purpose contained onomastic keywords that were unique to the project, the specific contents were vague to outside readers. In one case, no personnel were responsible for the appropriateness of the PDM. There were also some cases in which the terminal evaluation team formulated the PDMe by trimming the original PDM goals or setting up new indicators to make the project easier to evaluate.

Quality control of the initially prepared PDM and clarification of responsibility are required to make full use of the the principles for good PDM described above. The director of the overseas office, for example, should be responsible for the initially prepared PDM and the subsequent revision of the PDM. Furthermore, some crucial points must be clearly stipulated. Specifically, it is necessary to identify whether or not the initial PDM should be revised. If revised, it is essential to explain how the relevance, appropriateness, subsequent changes in input, and effectiveness are evaluated, how the PDM of a project under implementation and the revised PDM are treated, and the nature of monitoring and evaluation activities involved in the revised PDM.

■ Utilization of Development Objectives Chart

In order to complement the PDM at the ex-ante evaluation stage, it is necessary to utilize the development objectives chart. In JICA projects in the past, the relationships between a given project and its overall goal used to be presented in the project summary of PDM. However, as program approach and issue-specific approach progress, the number of projects is gradually growing in which not only JICA but also other donors are involved in several projects, thus constituting a large program. JICA projects are carried out as a part of the large program. In such cases, there exist some intermediate targets that come between the project purpose and overall goal. With the project purpose and overall goal shown in the PDM only, the causal relationships may not be fully explained. This can be a limit of presenting the outline of the project in the form of PDM.

In fact, this year's secondary evaluation observed a large discrepancy between project purposes and overall goals. For exam-

ple, in order to contain the word "poverty reduction," some of the overall goals included what is far from the causal relationships with project purpose or in some cases a pilot project was expected to expand as a full range national project in the overall goal. When the achievement of the project itself is measured by output, the relevance of a project is determined by project purpose and overall goal; and therefore, there is a concern that the relevance of the project may be questioned if the overall goal is out of reach. However, based on the recognition that the discrepancy between project purposes and overall goals would be resolved by coordination with other projects of JICA, other projects of the partner countries and other donors, some projects set much higher overall goals than project purposes at the time of project formulation.

In such a case, the relation between and the positions of a given project and other projects should be demonstrated by using development objectives charts from the time of ex-ante evaluation to confirm whether the expected outputs and project purposes are appropriately explained. It might also be necessary to include in the external factors of PDM such descriptions as "project A will be carried out by XXX as planned and outcomes will be obtained," etc. It is necessary to carry out terminal evaluation while confirming the progress of other assistance concerned, in addition to data collection concerning the project.

(6) Summary of Project Evaluation by the Secondary Evaluators based on the Reports

"Relevance" of the target projects was generally high and "effectiveness" generated good outcomes as a whole although there are discrepancies between projects. "Efficiency," "impact," and "sustainability" also achieved a certain level.

When "five evaluation criteria" were examined in terms of viewpoints, the difference among viewpoints was large for "relevance." The validity of project implementation was high in terms of consistency with Japan's aid policies, JICA Country Programs, and development policies of the partner countries, and the adequacy of the implementation as ODA. On the other hand, the viewpoint of the appropriateness of project design was rated relatively low as to whether the approach was appropriate as an effective solution to the development issues, whether the selection of target areas or target groups was appropriate, and whether Japanese technology was superior. With regard to sustainability, rated high was the viewpoint as to whether the positioning of activities in the policies and organization of the implementing agency were stable enough to produce continuous effects after the termination of cooperation. However, the viewpoints as to whether the budgets, including operating expenses, were secured and whether measures to ensure budgets were sufficient was below 3.0 on average.

Despite some problems described above, when evaluation results were chronologically compared from fiscal 2002 to fiscal 2004, there was a significant difference between fiscal 2002/2003 and fiscal 2004 though little difference between fiscal 2002 and fiscal 2003. Not all projects in fiscal 2004 were analyzed this time, and there is some reservation that a large portion of the selected projects may have been relatively good ones whose evaluation reports could be complied easily at earlier dates. Yet, it is natural to interpret that the quality of projects has improved.

When good projects and poor projects were compared, there was a significant difference in each of the five evaluation criteria. Since the difference is large in effectiveness, sustainability, and efficiency, it is necessary to pay particular attention to these criteria when managing projects.

The evaluation results of projects as described above and the quality of primary evaluation exhibit a certain relation. High quality of project indicates that the initial plans are appropriately designed, necessary data are accumulated through periodical monitoring, and a high quality report is easily formulated. On the other hand, if the project evaluation is low, the variance is large. Good projects in general tend to have high quality terminal evaluation; however, regardless of the quality of projects, it is necessary, when evaluating any project, to collect information from various sources, conduct assessment and analysis of performance objectively and accurately, and offer a clear description using tables and figures.

Appendix 1 List of Projects Subject to Secondary Evaluation in Fiscal 2005

Fiscal 2002 (Targets of Secondary Evaluation in Fiscal 2004)		
Social Development	Malaysia	Japan-Malaysia Technical Institute: JMTI
Social Development	Thailand	Development of the Method of Urban Development
Social Development	Brazil	The Urban Transport Human Resources Development Project
Social Development	Paraguay	Japan-Paraguay Skill Development Promotion Center
Health and Medical Care	Jordan	The Project for Family Planning and Gender in Development (Phase 2)
Health and Medical Care	Kenya	Kenya Medical Training College Project
Forest and Natural Environment	Malaysia	The Project for the Aquatic Resource and Environmental Studies of the Straits of Malacca in UPM
Forest and Natural Environment	Bolivia	The Afforestation and Erosion Control Project in the Valley of Tarija
Forest and Natural Environment	Uruguay	Forest Products Testing Project
Forest and Natural Environment	Madagascar	The Aquaculture Development Project in the Northwest Coastal Region of Madagascar
Mining and Industrial Development	Laos	The Project on Electric Power Technical Standard Establishment
Fiscal 2003 (Targets of Secondary Evaluation in Fiscal 2004)		
Social Development	Indonesia	Regional Development Policies for Local Government
Social Development	Philippines	The Cebu Socio-economic Empowerment and Development Project
Health and Medical Care	Ghana	The Infectious Diseases Project at the Noguchi Memorial Institute for Medical Research
Agricultural Development	Myanmar	Irrigation Technology Center Project (Phase 2)
Agricultural Development	Thailand	The Modernization of Water Management System Project
Agricultural Development	El Salvador	The Project for the Strengthening of Agricultural Technology Development and Transfer
Agricultural Development	Mexico	The Agricultural Machinery Test and Evaluation Project
Forest and Natural Environment	Laos	The Aquaculture Improvement and Extension Project
Forest and Natural Environment	Brazil	Brazilian Amazon Forest Research Project (Phase 2)
Mining and Industrial Development	Viet Nam	Modernization of Industrial Property Administration Project
Fiscal 2003 (New Targets)		
Social Development	Indonesia	Detailed Design of Flood Control and Water Resources Development Project in Semarang
Social Development	Philippines	Technology Development of Electronic Navigational Charts
Social Development	Paraguay	Japan-Paraguay Skill Development Promotion Center (Extended)
Social Development	Senegal	High-level Technician (BTS) Training Project at the Senegal-Japan Vocational Training Center
Social Development	Tanzania	Sokoine University of Agriculture Center for Sustainable Rural Development: SCSRD
Social Development	Uganda	Nakawa Vocational Training Institute Project (Extended)
Social Development	Zambia	Technical and Vocational Training Improvement Project (Aftercare)
Health and Medical Care	China	Anhui Primary Health Care Technical Training Center
Health and Medical Care	Indonesia	Malaria Control in Lombok and Sumbawa Islands
Health and Medical Care	Malaysia	The Project for Strengthening of Food Safety Programme
Health and Medical Care	Thailand	Project for Strengthening of National Institute of Health Capabilities for Research and Development on AIDS and Emerging Infectious Diseases
Health and Medical Care	Dominican Republic	Medical Education and Training Project
Health and Medical Care	Mexico	Reproductive Health – Prevention of Uterine Cervical Cancer
Health and Medical Care	Tunisia	The Project for Strengthening of Reproductive Health Education
Health and Medical Care	Yemen	The Tuberculosis Control Project (Phase 3)
Health and Medical Care	Ethiopia	Laboratory Support for Polio Eradication: LAST Polio Project
Health and Medical Care	Madagascar	Project for the Global Improvement for the Mahajanga University Hospital Center
Agricultural Development	Cambodia	Improvement of the Survey and Forecast System on Meteorology and Agro-meteorology
Agricultural Development	China	Enhancement of Agricultural Extension System Project
Agricultural Development	Malaysia	The Project for the Development of Technology Related to the Processing of Feed Based on Agro-industrial By-products of Oil Palms Production (Follow-up)
Agricultural Development	Argentina	The Joint Study on Biological Control of Soil-borne Plant Diseases
Agricultural Development	Brazil	The Technological Development Project for Sustainable Agriculture in Eastern Amazonia
Agricultural Development	Iran	The Project of Haraz Agricultural Human Resources Development Center
Agricultural Development	Kenya	Promotion of Sustainable Community Based Small-holder Irrigation
Forest and Natural Environment	Indonesia	The Mangrove Information Center Project
Forest and Natural Environment	Nepal	Community Development and Forest/Watershed Conservation Project (Phase 2)
Forest and Natural Environment	El Salvador	The Project on the Aquaculture Development
Mining and Industrial Development	Indonesia	Project on Supporting Industries Development for Casting Technology
Fiscal 2004 (New Targets)		
Social Development	Viet Nam	Project on the Improvement of Higher Maritime Education
Social Development	Viet Nam	The Project for Strengthening Training Capacity for Technical Workers in the Hanoi Industrial College
Social Development	Turkey	The Project on Improvement of Maritime Education
Health and Medical Care	Cambodia	The Maternal and Child Health Project (Phase 2)
Health and Medical Care	Laos	The Project for the Improvement of Sethathirath Hospital
Health and Medical Care	Thailand	The Project for the Asian Center for International Parasite Control
Health and Medical Care	Viet Nam	The Bach Mai Hospital Project for Functional Enhancement
Health and Medical Care	Nicaragua	The Project for Strengthening of the Local System of Integrated Health Care (SILAIS) of Granada
Agricultural Development	Malaysia	Molecular Characterization of NIPAH Virus in Animals
Agricultural Development	Philippines	Environmental and Productivity Management of Marginal Soils
Agricultural Development	Chile	Improvement of Productivity for the Small-scale Dairy Farmers Project
Forest and Natural Environment	China	The Model Afforestation Project in Sichuan
Forest and Natural Environment	China	Research Project on Timber from Man-made Forests
Forest and Natural Environment	Philippines	Environmental and Productivity Management of Marginal Soils
Forest and Natural Environment	Thailand	The Reforestation and Extension Project in the Northeast of Thailand (Phase 2)
Mining and Industrial Development	Thailand	The Project on the Industrial Water Technology Institute (Phase 2)
Mining and Industrial Development	Argentina	Regional Geologic Mapping with Advanced Satellite Sensors

Appendix 2

Secondary Evaluation Check Sheet

Rating criteria

1) Rate viewpoints and criteria in green cells based on a scale of 1 to 5. [I – III]
5: Sufficient/high
4: Fairly sufficient/high
3: Average
2: Slightly insufficient/low
1: Insufficient/low
*: Cannot tell

2) Rate familiarity in green cells choosing from the dropdown list.

3) Write down highlights and notable points (including good practices) in the space for comment. [I – IV]

I. Evaluability

1. Evaluability of the Initially Prepared Project Design Matrix (PDM)		Rating
Viewpoint	The initially designed PDM is usable as an evaluation framework without significant changes in its objectives and indicators.	
2. Evaluability of Outputs, Project Purpose and Overall Goal		Rating
Viewpoint	The indicators are clearly defined for each output, project purpose, and overall goal, with specific target values and beneficiaries. They can be used to measure the level of the project achievement.	
3. Logic of Project Design		Rating
Viewpoint	The PDM for the evaluation describes a clear and realistic logic flow from Overall Goal - Project Purpose - Outputs - Inputs, considering important external assumptions.	
4. Project Monitoring		Rating
Viewpoint	Monitoring of outputs, activities, and inputs was regularly conducted, and the information including statistical data was accumulated during project implementation.	
Comment		Overall

II. Key Evaluation Criteria

1 Evaluation Framework		
1. Time Frame of Evaluation Study		Rating
Viewpoint	Necessary field survey activities such as data collection and discussion with counterparts are appropriately set within the time frame of the evaluation study. Time frame also contains preparations such as distribution of questionnaires, and is appropriate in terms of timing, length, and schedule of the evaluation study.	
2. Evaluation Team Composition– Impartiality and Specialty		Rating
Viewpoint	The evaluation team members are selected on an impartial basis and with balanced specialty.	
3. Level of Counterpart Participation		Rating
Viewpoint	The counterparts understand evaluation process, and share responsibilities for evaluation activities with JICA.	
Comment		Overall

2 Data Collection		
1. Evaluation Questions		Rating
Viewpoint	Evaluation questions are in line with evaluation purposes and set properly in the evaluation grid. General questions as to the five evaluation criteria are narrowed down to more specific sub-questions to identify necessary information/data to be collected.	
2. Appropriateness of Data Collection Methods and Data Sources		Rating
Viewpoint	Several different data collection methods are used to increase accuracy and reliability of the data/information obtained. The data/information is obtained from a broad range of stakeholders, including the end beneficiary groups to limit bias of the data collected.	
3. Data/Information Sources		Rating
Viewpoint	The sources of the data/information are adequately explained in the evaluation report.	
4. Sufficiency of Data/Information Obtained		Rating
Viewpoint	Data collection is conducted based on the evaluation grid, and the data/information was sufficient to answer the evaluation questions, and additional information/data is gathered for unexpected and newly confronted questions during the evaluation process.	
Comment		Overall
3. Analysis/ Evaluation		
3.1 Assessment of Performance		
1. Measurement of Results		Rating
Viewpoint	Achievement level of outputs, project purpose, and overall goal are measured quantitatively or/and qualitatively against the target values set by the indicators.	
2. Examination of Project Implementation Process		Rating
Viewpoint	The project implementation process is thoroughly examined, through which impeding and/or promoting factors to achievement of outputs, project purpose, and overall goal are identified.	
3. Examination of Causal Relationships—Logic of Project Design [1]		Rating
Viewpoint	The logic of project design is thoroughly verified, through which impeding and/or promoting factors to achievement of outputs, project purpose, and overall goal are identified.	
4. Examination of Causal Relationships—Before and After [2]		Rating
Viewpoint	The causal relationships are thoroughly examined to verify that effects for the beneficiaries have resulted from the project interventions.	
Comment		Overall
3.2 Analysis		
1. Objectivity of Analysis		Rating
Viewpoint	The data is objectively analyzed, based on a series of scientific discussions, and an effort is made to quantify the data where feasible.	
2. Holistic Analysis		Rating
Viewpoint	The data interpretation is drawn by examination and analysis of different methods, and from various aspects.	
3. Analysis of Promoting and Impeding Factors		Rating
Viewpoint	Factors that promote and impede effects are adequately analyzed in light of the project logic (cause-effect) and the project implementation process (such as project management).	
Comment		Overall

3.3 Evaluation by Five Criteria		
1. Relevance		Rating
Viewpoint	Perspectives for evaluation of "Relevance" (validity and necessity of a project in light of needs of beneficiaries, project implementation as an appropriate approach to problem solving, consistency of policies, etc.) are sufficiently covered.	
2. Effectiveness		Rating
Viewpoint	Perspectives for evaluation of "Effectiveness" (achievement level of project purpose, causal relationships between outputs and project purpose, etc.) are sufficiently covered.	
3. Efficiency		Rating
Viewpoint	Perspectives for evaluation of "Efficiency" (comparison with other similar projects through cost analysis, cost-effectiveness analysis, etc.) are sufficiently covered.	
4. Impact		Rating
Viewpoint	Perspectives for evaluation of "Impact" (achievement level of overall goal, causal relationships between project purpose and overall goal) are sufficiently covered.	
5. Sustainability		Rating
Viewpoint	Perspective for evaluation of "Sustainability" (probability of effects to be continued and outcomes to be produced in terms of policies and systems, organizational and financial aspects, technical aspects, socio-culture, and environment) are sufficiently covered.	
6. Conclusion		Rating
Viewpoint	The conclusion is drawn based on holistic viewpoints on the basis of the five evaluation criteria.	
Comment		Overall
4. Recommendations/Lessons Learned		
1. Relevance and Credibility of Recommendations		Rating
Viewpoint	The recommendations are based on the information obtained through the process of data analysis and interpretation. As a result, the recommendations are objective and convincing.	
2. Sufficiency of Recommendations		Rating
Viewpoint	The recommendations consider all the impeding/promoting factors identified during the evaluation process.	
3. Usability of Recommendations		Rating
Viewpoint	The recommendations are practical and useful for feedback and follow-ups, with a specific time frame.	
4. Relevance and Credibility of Lessons Learned		Rating
Viewpoint	The lessons learned are based on the information obtained through the process of data analysis and interpretation. As a result, the lessons learned are objective and convincing.	
5. Sufficiency of Lessons Learned		Rating
Viewpoint	The lessons learned consider all the impeding/promoting factors identified during the evaluation process.	
6. Usability of Lessons Learned		Rating
Viewpoint	The lessons are generalized and conceptualized so that they are widely applicable.	
Comment		Overall
5. Reporting		
1. Presentation/Legibility and Clarity		Rating
Viewpoint	The evaluation report is simple and clear, and understandable to readers—in light of the structure, font, terminology, and data presentation. Logical structure and major points are clearly described in an easily understandable manner.	
2. Utilization of Tables and Figures		Rating
Viewpoint	Tables and figures are effectively utilized to visually present statistics and analysis results.	
3. Presentation of Primary Data		Rating
Viewpoint	Sufficient primary data such as those on targets and results of interviews and questionnaires are presented properly in the report.	
Comment		Overall

III. Evaluation of the Project Based on the Report

Fill in comments if there are any external important assumptions that might affect the following Five Evaluation Criteria.

1. Relevance		
1. Validity		Rating
Viewpoint	The project is consistent with Japan's aid policies, JICA Country Program, and development policies of the partner country. Its implementation by means of ODA is relevant.	
2. Necessity		Rating
Viewpoint	The project is in line with the needs of the target group, area, and society. Those needs are still present and logically understood including priority.	
3. Appropriate Approach		Rating
Viewpoint	The approach is appropriate to solve the preset development issue (overall goal). The selection of target area and group is appropriate. Japanese technology is superior.	
Comment		Overall
2. Effectiveness		
1. Achievement Level of Project Purpose		Rating
Viewpoint	Project purpose has been (is going to be) achieved.	
2. Causal Relationships between Outputs and Project Purpose		Rating
Viewpoint	Project purpose has been (is going to be) achieved as a result of outputs. Important assumptions which might affect the achievement of outputs and project purpose were properly identified. There were special factors which inhibited or promoted effectiveness.	
Comment		Overall
3. Efficiency		
1. Cost-effectiveness		Rating
Viewpoint	Efforts to cut down on costs were made (using local resources).There was no alternative means that could have led to the same achievements at lower costs. It was impossible to produce greater achievements at the same costs. Compared to similar projects of other donors and the partner country, the cost-effectiveness was high.	
2. Appropriate Implementation Process		Rating
Viewpoint	The inputs were made in a timely manner with appropriate scale and quality.	
Comment		Overall
4. Impact		
1. Achievement Level of Overall Goal		Rating
Viewpoint	Effects planned in the project (overall goal) have been achieved as a result of achievement of project purpose. Problem-solving for the target project has progressed.	
2. Causal Relationships between Project Purpose and Overall Goal		Rating
Viewpoint	Impact was generated as a result of achievement of project purpose. There were special factors that promoted or impeded planned effects including important assumptions.	
3. Unintended Positive and Negative Impact		Rating
Viewpoint	There are political impacts and economical impacts on the target society, inside the implementing agency, and on the beneficiary. Other impacts on organization, development of related regulation and laws, gender equality, human rights, disparity between rich and poor, peace and war, environmental protection are present. There are special factors that brought unintended positive and negative impacts.	
Comment		Overall

5. Sustainability		
1. Mechanism of Securing Sustainability		Rating
Viewpoint	Mechanisms and devices for securing sustainability (management capacity of the implementing agency, policy support from the supervising agency, demand for activities of the implementing agency, securing financial basis) were considered in the project.	
2. Level of Sustainability		Rating
Viewpoint	Effects aimed for in the project (project purpose and overall goal) are (will be) sustained after the termination of cooperation.	
3. Organizational Sustainability		Rating
Viewpoint	The positioning of activities in the policies and organization of the implementing agency is stable enough to conduct activities that will continue effects after the termination of cooperation.	
4. Technological Sustainability		Rating
Viewpoint	Technology and capacity acquired in the project are maintained and expanded. Equipment is properly maintained and managed.	
5. Financial Sustainability		Rating
Viewpoint	Budget including operating expenses is secured. Measures for securing budget are sufficient.	
Comment		Overall

IV. Overall Comment

V. Familiarity

1. Prior Information about the Project		Rating
Viewpoint	1. None 2. Know by name 3. Know some 4. Know well (have read reports, etc.) 5. Know very well (have conducted study, etc.)	
2. Familiarity with Area		Rating
Viewpoint	1. None 2. Know by name 3. Know some 4. Know well (have read reports, etc) 5. Know very well (have conducted study, etc.)	
3. Familiarity with Specialty		Rating
Viewpoint	1. None 2. Know by name 3. Know some 4. Know well (have read reports, etc) 5. Know very well (have conducted study, etc.)	