Part 4

Secondary Evaluation by the Advisory Committee on Evaluation

Secondary Evaluation by the Advisory Committee on Evaluation

JICA established the Advisory Committee on Evaluation in fiscal 2002. The objectives of the committee are to improve the evaluation system and methods based on advice from external experts and to increase objectivity by having them examine evaluation results. The committee members are external intellectuals such as academics, NGOs, concerned parties from international organizations, and journalists, all of whom have knowledge and experience in development assistance and evaluation. The members discuss various agenda including issues involved in enhancing evaluation systems and promoting feedback of evaluation results. They also conduct secondary evaluation of JICA's evaluations and present recommendations on improving evaluation methods and projects through evaluation.

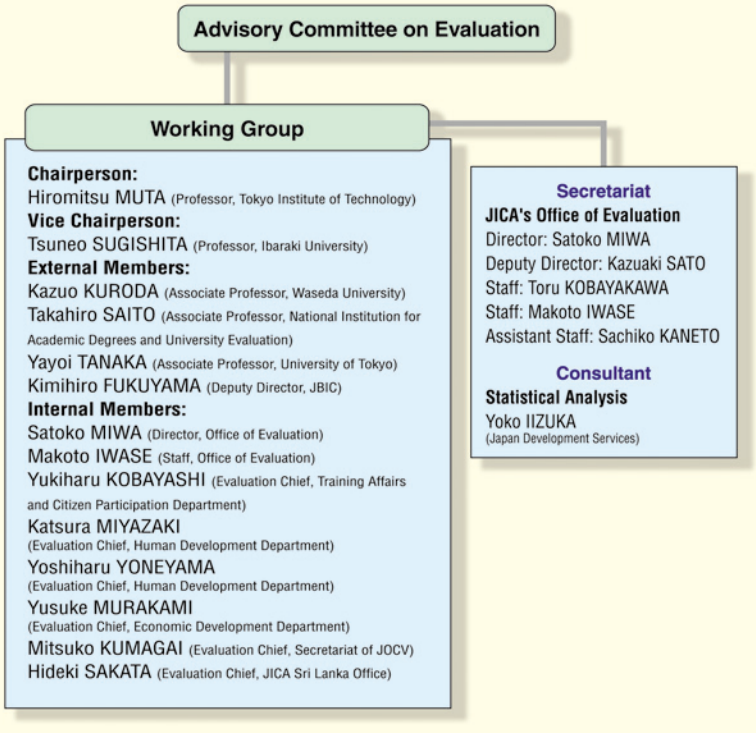
The first secondary evaluation was conducted in fiscal 2002 and 2003 targeting 40 terminal evaluations carried out in fiscal 2001. The results were presented in the JICA Annual Evaluation Report 2003. JICA conducts, with the participation of external specialists and concerned parties in the partner country, terminal evaluation on a project as an internal evaluation to find out if the outputs and outcomes have been successfully generated as planned by the project, and to arrive at necessary and responsible decision-making for better project management. However, since JICA and related organizations in the partner country are responsible parties for implementing the project itself, it is sometimes pointed out that internal evaluations led by those parties may lack neutrality or objectivity. To address this problem, a secondary evaluation is conducted to increase the transparency and objectivity of JICA's internal evaluations.

It is also one of the meaningful aspects of the secondary evaluation that evaluations carried out by JICA are assessed from a third-party perspective and the current conditions and problems involved in the internal evaluation are clarified to improve its methods and quality. The first secondary evaluation was granted with a number of valuable suggestions and advice on planning and implementing procedures for evaluation study, utilization of evaluation results, operation and management of projects, etc. The results of the above-mentioned first evaluation were reflected in the revision of the JICA project evaluation guidelines in fiscal 2004 and JICA is currently making efforts to consolidate evaluation practices based on the guidelines.

The results of the second secondary evaluation that was conducted in fiscal 2004 are contained in the Annual Evaluation Report 2004. Secondary evaluation examines the quality of JICA's internal evaluations and outcomes of JICA's projects observed from evaluation reports. Based on the effectiveness of secondary evaluation, which was recognized as a result of the first secondary evaluation, deeper analysis was carried out this time on current conditions and problems regarding JICA's evaluations, and at the same time year-to-year comparisons of the changes in evaluation quality as well as differences in perspectives between external and internal evaluators were examined. For this purpose, the Secondary Evaluation Working Group comprising external and internal evaluators was established under the Advisory Committee on Evaluation, which is illustrated in Figure 4-1.

JICA is devoted to increasing the transparency and improving the quality of evaluations, using the secondary evaluation results presented in the following chapter.

Figure 4-1 Secondary Evaluation System of Fiscal 2004



Chapter 1 Results of Secondary Evaluation Fiscal 2004

—Improving the Objectivity and Quality of Evaluation

Advisory Committee on Evaluation/ Secondary Evaluation Working Group

1-1 Objectives, Targets, Methods of Evaluation

(1) Objectives

The significance of secondary evaluation conducted by external experts is found in ensuring the transparency of internal evaluations and the credibility of evaluation results.

JICA conducts evaluations on JICA's projects internally. External experts, consultants, and concerned parties of the partner country usually participate in the evaluation processes; however, they are often involved with the projects in such ways as being related to supporting organizations in Japan and implementing bodies of the partner country. Participation of parties who are familiar with a target project is an advantage that allows for detailed evaluations. However, at the same time, the possibility that it may arouse problems regarding the neutrality and objectivity of evaluation cannot be denied. To address this, secondary evaluation by external experts is useful for increasing transparency and reviewing the appropriateness of internal evaluations.

Each evaluator exhibits some personal patterns of evaluation that affect evaluation results. Though external evaluations can be objective in that evaluators have no interests in the projects, they are not necessarily superior to internal evaluations in terms of neutrality and impartiality as long as there is the possibility of personal bias in evaluation. This was made clear in the first secondary evaluation of fiscal 2003 conducted by the Advisory Committee on Evaluation. Therefore, in order to obtain more universal and reliable evaluation results, it is necessary to reach conclusions that are free from personal bias in evaluation by taking several evaluators' perspectives. However, it is not practical in terms of cost to have so many external evaluators evaluate a project or to dispatch another study team. Although secondary evaluation is somewhat limited in that it is based on the results of primary evaluations, it is expected that less biased and highly reliable evaluation results can be obtained through the participation of many secondary evaluators different from the primary evaluators.

In addition to the above-mentioned points, secondary evaluation is significant in that it contributes to improvements in the quality of primary evaluations. Evaluation by third parties is useful to objectively verify the appropriateness of the evaluation framework, methods, analysis, value judgment, and the way of presenting the results of primary evaluation, etc. Evaluation by third parties is also useful in that it reveals challenges for improving the quality of evaluation through this verification process. The first secondary evaluation proved the effectiveness of secondary evaluation in this respect, in addition to effectiveness in increasing transparency and credibility of evaluation. The improvement in the quality of evaluations is indispensable so that evaluations could take roles in improving projects as well as for securing accountability. Contribution to the improvement of primary evaluation is, indeed, the essential significance of secondary evaluation.

Based on the significance described above, the secondary evaluation of fiscal 2004 was performed to answer the following evaluation questions.

Examination of the quality of primary evaluation (internal evaluation)

a. Does the evaluation satisfy a certain quality? What has to be done to achieve higher quality?
b. Has the quality of evaluations improved every year?
c. What are the differences in perspectives between internal and external evaluations that need to be considered in order to improve the quality of evaluations?

Examination of the outcomes of target projects

a. Was the target project effective and efficient from the perspectives of the secondary evaluators?
b. What are the factors that influence the effectiveness and efficiency of the target project?
c. Are there differences in the results of the secondary evaluation on the outcomes of the target project between internal evaluations and external evaluations?

Most of JICA's evaluations are mainly conducted internally, as already mentioned. In order to secure neutrality and objectivity, it is necessary to incorporate external views such as those of secondary evaluations. Nonetheless, it is still essential to perform internal evaluations with as much objectivity and credibility as possible. In order to extract important points to consider in achieving this, we will examine the differences between internal evaluations and external evaluations on the quality of evaluations and the outcomes of target projects.

(2) Evaluators

In view of these objectives, the secondary evaluation of fiscal 2004 set up a Secondary Evaluation Working Group with external and internal evaluators under the Advisory Committee on Evaluation, and evaluations were conducted by the working group accordingly. The external evaluators in the working group are two members of the Advisory Committee on Evaluation, three external members recommended by the Japan Evaluation Society, and one staff member from the Japan Bank for International Cooperation, which is another implementing body of ODA programs like JICA. The internal evaluators are two staff members from the Office of Evaluation, Planning and Coordination Department, which supervises the quality of evaluations within JICA, and six evaluation chiefs from the project implementation departments.

Evaluators of the Secondary Evaluation

Chairperson of the Secondary Evaluation Working Group

Hiromitsu MUTA:
Professor, Director of the Center for Research and Development of Educational Technology, Tokyo Institute of Technology (Chairperson of the Advisory Committee on Evaluation)

Vice Chairperson:

Tsuneo SUGISHITA:
Professor, College of Humanities, Ibaraki University, Formerly with Yomiuri Shinbun (Member of the Advisory Committee on Evaluation)

External Members:

Kazuo KURODA:
Associate Professor, Graduate School of Asia Pacific Studies, Waseda University (recommended by the Japan Evaluation Society)

Takahiro SAITO:
Associate Professor, National Institution for Academic Degrees and University Evaluation (recommended by the Japan Evaluation Society)

Yayoi TANAKA:
Associate Professor, Department of Civil Engineering, University of Tokyo (recommended by the Japan Evaluation Society)

Kimihiro FUKUYAMA:
Deputy Director, Development Assistance Operations Evaluation Office, Project Development Department, Japan Bank for International Cooperation

Internal Members:

Satoko MIWA:
Director, Office of Evaluation, Planning and Coordination Department, JICA

Makoto IWASE:
Staff, Office of Evaluation, Planning and Coordination Department, JICA

Yukiharu KOBAYASHI:
Team Leader, Office of Citizen Participation, Training Affairs and Citizen Participation Department, JICA (Evaluation Chief)

Katsura MIYAZAKI:
Team Leader, Group III (Health I), Human Development Department, JICA (Evaluation Chief)

Yoshiharu YONEYAMA:
Team Leader, Group IV (Health II), Human Development Department, JICA (Evaluation Chief)

Yusuke MURAKAMI:
Team Leader, Group II (Natural Resources and Energy), Economic Development Department (Evaluation Chief)

Mitsuko KUMAGAI:
Team Leader, Administration and Planning Group, Secretariat of JOCV (Evaluation Chief)

Hideki SAKATA:
Deputy Director, JICA Sri Lanka Office (Evaluation Chief)

* External evaluators: Chairperson and Vice Chairperson of Secondary Evaluation Working Group, external members
Internal evaluators: internal members (Positions were effective as of October 2004.)

(3) Evaluation Targets

This secondary evaluation targeted 38 terminal evaluations conducted in fiscal 2002 and 10 terminal evaluations in fiscal 2003 whose reports had been disclosed at the time of the launch of the secondary evaluation at the beginning of 2004. The evaluation results for fiscal 2001, which had already been evaluated by the first secondary evaluation, were analyzed again in order to perform a year-to-year comparison. For this comparison analysis, 10 reports with evenly distributed scores from high to low were selected based on the evaluation scores of quality conducted by the first secondary evaluation. All of the 58 primary evaluation results that were subject to the secondary evaluation are listed in Table 4-1.

(4) Evaluation Methods

Evaluators read the primary evaluation reports and graded them using the secondary evaluation check sheet shown in Table 4-2.

1) Evaluation Viewpoints

The check sheet consists of criteria to be used to examine the quality of primary evaluations and criteria to be used to examine the outcomes of the project subject to primary evaluation.

a. Examination of the quality of primary evaluation

As for the examination of the quality of primary evaluation, evaluation criteria are determined by taking into consideration (1) key evaluation criteria, (2) general criteria for good evaluation, and (3) preconditions for conducting appropriate evaluation (evaluability of target projects). Evaluation viewpoints for each criterion are as follow.

- Key evaluation criteria:
  The appropriateness of evaluation methods is evaluated according to the following criteria: framework of evaluation study, methods of information collection, methods of analysis and evaluation (assessment of performance, analysis, and DAC Five Evaluation Criteria), extraction and presentation of recommendations and lessons, and compilation of reports.
- General criteria for good evaluation:
  The quality of evaluation is comprehensively assessed according to the criteria for good evaluation provided in the JICA Evaluation Guidelines: usefulness of evaluated information, impartiality and independence, credibility, and participation of the partner country.
- Evaluability:
  This criterion assesses whether a project has been adequately planned and implemented for subsequent verification of the results, such as the feasibility of verifying achievement of purposes and outputs, and reasonableness of the logic model behind the project. In addition, it is also checked whether monitoring data necessary for evaluation

Table 4-1 List of Projects Subject to Secondary Evaluation

FY 2001 (only for the analysis of year-to-year changes in quality)		
1	China	The Project for the Beijing Municipal Education and Training Center for Fire Fighting and Prevention
2	India	The Project for Promotion of Popularizing the Practical Bivoltine Sericulture Technology
3	Indonesia	Dairy Technology Improvement Project in Indonesia
4	Indonesia	Project for Improvement of Agricultural Extension and Training System
5	Malaysia	The Project on Risk Management of Hazardous Chemical Substances
6	Philippines	Research and Development Project on High Productivity Rice Technology
7	Morocco	Upgrading Exploration Technology of Mineral Resources in the Kingdom of Morocco
8	Argentina	The Mine Pollution Control Research Center in the Argentine Republic
9	Brazil	The Research Project on Small-scale Horticulture in Southern Brazil
10	Brazil	The Clinical Research Project of State University of Campinas in Brazil
FY 2002		
1	Bangladesh	The Poultry Management Techniques Improvement Project in the People's Republic of Bangladesh
2	Cambodia	Secondary School Teacher Training Project in Science and Mathematics
3	India	The Project for Prevention of Emerging Diarrheal Diseases in India
4	Indonesia	The Project for the National Vocational Rehabilitation Center for Disabled People
5	Indonesia	The Project for Development of Science and Mathematics Teaching for Primary and Secondary Education
6	Indonesia	Technical Cooperation Project for Ensuring the Quality of MCH Services through MCF Handbook
7	Indonesia	Development of High Quality Seed Potato Multiplication System Project
8	Indonesia	The Forest Tree Improvement Project (Phase 2) in the Republic of Indonesia
9	Indonesia	Biodiversity Conservation Project (Phase 2)
10	Laos	The Agricultural and Rural Development Project in Vientiane Province in the Lao People's Democratic Republic (Phase 2)
11	Laos	The Forest Conservation and Afforestation Project (Phase 2) in Lao People's Democratic Republic
12	Laos	The Project on Electric Power Technical Standard Establishment in Lao People's Democratic Republic
13	Malaysia	Japan-Malaysia Technical Institute: JMTI
14	Malaysia	The Project for the Aquatic Resource and Environmental Studies of the Straits of Malacca in UPM
15	Philippines	The Project for Upgrading Human Resource Development for Air Navigation Systems Specialist at the Civil Aviation Training Center Manila
16	Philippines	The Project for Enhancement of Capabilities in Flood Control and Sabo Engineering of the Department of Public Works and Highways
17	Philippines	The Project on Electrical and Electronics Appliances Testing in the Republic of Philippines
18	Philippines	Modernization of Industrial Property Administration
19	Sri Lanka	Dental Education Project at University of Peradeniya in Sri Lanka
20	Thailand	The Research Center for Communication and Information Technology (ReCCIT), King Mongkut's Institute of Technology, Ladkrabang, (KMITL), the Kingdom of Thailand
21	Thailand	Development of the Method of Urban Development
22	Thailand	Project for Model Development of Comprehensive HIV/AIDS Prevention and Care
23	Viet Nam	The Education and Research Capability Building Project of Hanoi Agricultural University
24	Jordan	The Project for Family Planning and Gender in Development (Phase 2)
25	Jordan	Information Technology Upgrading Project
26	Turkey	The Infectious Diseases Control Project in the Republic of Turkey
27	Ethiopia	The Groundwater Development and Water Supply Training Project
28	Kenya	Strengthening of Mathematics and Science in Secondary Education
29	Kenya	Kenya Medical Training College Project
30	Madagascar	The Aquaculture Development Project in the Northwest Coastal Region of Madagascar
31	Bolivia	The Afforestation and Erosion Control Project in the Valley of Tarija in Bolivia
32	Brazil	The Urban Transport Human Resources Development Project
33	Jamaica	The Project for Strengthening of Health Care in the Southern Region
34	Panama	The Cattle Productivity Improvement Project in the Republic of Panama
35	Paraguay	Japan-Paraguay Skill Development Promotion Center
36	Paraguay	Project on Upgrading Verification and Inspection Technology in the Area of Mass
37	Uruguay	Forest Products Testing Project in Uruguay
38	Micronesia	The Fisheries Training Project in Federated States of Micronesia
FY 2003		
1	Indonesia	Regional Development Policies for Local Government
2	Laos	The Aquaculture Improvement and Extension Project
3	Myanmar	Irrigation Technology Center Project (Phase 2)
4	Philippines	The Cebu Socio-economic Empowerment and Development Project
5	Thailand	The Modernization of Water Management System Project in Thailand
6	Viet Nam	Modernization of Industrial Property Administration Project
7	Ghana	The Infectious Diseases Project at the Noguchi Memorial Institute for Medical Research
8	Brazil	Brazilian Amazon Forest Research Project (Phase 2)
9	El Salvador	The Project for the Strengthening of Agricultural Technology Development and Transfer in El Salvador
10	Mexico	The Agricultural Machinery Test and Evaluation Project in Mexico

have been adequately accumulated during the project implementation.

Evaluability was included because the first secondary evaluation pointed out that the original determinations of the project purpose, data collection methods, and logical framework were the issues associated with the quality of evaluation. Although examination of evaluability from the report alone naturally has limitations, this item was added on a trial basis this time for the purpose of further analyzing the current situation of projects' evaluability and the connection between evaluability and evaluation.

b. Examination of the outcomes of the project subject to primary evaluation

To evaluate the outcomes of the projects subject to primary evaluation, DAC Five Evaluation Criteria (relevance, effectiveness, efficiency, impact, and sustainability) and general evaluation were adopted.

2) Evaluation Scores

a. Examination of the quality of primary evaluation

The nine evaluation criteria (seven criteria based on key evaluation perspectives, general criteria for good evaluation, and evaluability) were examined on a scale of 10. Each crite-

rion provides space for comments to collect qualitative information that supplements quantitative ratings.

In addition, several sub-criteria are provided for specifying the viewpoints of each evaluation criterion. These sub-criteria were newly set to clarify at which angle the secondary evaluation is conducted and standardize evaluation views. The sub-criteria are based on the items to be considered when conducting evaluations, which are provided in the JICA Evaluation Guidelines. These sub-criteria were also graded on a three-level scale in the secondary evaluation.

However, this secondary evaluation did not clarify the weights of the sub-criteria; instead, each evaluator graded each evaluation criterion independently from ratings of its sub-criteria. This is because it was decided that the working group would come up with more appropriate weightings, which should be applied to secondary evaluation in the future as a result of statistical analysis over correlation between each criterion and its sub-criteria, instead of applying discretional weightings in advance.

No total scores are added up for each criterion. This is because a high total score does not always mean good evaluation when, for example, a high total score may contain extremely low scores for some criteria.

b. Examination of the outcomes of the project subject to primary evaluation

Each of the six evaluation criteria (DAC Five Evaluation Criteria and overall evaluation) was graded on a scale of 10. In the same way that the quality of primary evaluation was examined, each criterion provides space for comments to collect qualitative information that supplements quantitative ratings.

c. Methodology for rating

When primary evaluations are rated by all the evaluators after they read all the reports and the mean of the scores obtained is calculated, the opinions of all the members can be reflected and the final score can be free of personal evaluation biases. However, this method is impractical due to the enormous workload of each evaluator. On the contrary, if the number of evaluators for each report is limited, the average score is greatly affected by the personal evaluation biases of the evaluators in charge, thus generating biased evaluation results. In order to solve this problem and obtain more universal scores, the secondary evaluation of fiscal 2004 devised methods of statistical analysis in the same manner as the first secondary evaluation. This was designed to calculate mean values that are estimated on the assumption that every evaluator had read all the reports.

In other words, the scores of each evaluator can theoretically be divided into two parts: the part that is free of personal evaluation bias of the evaluator (true score of the evaluation target), and the part that is affected by the personal evaluation bias of the evaluator (coefficient of evaluation tendencies). Therefore, if both parts are separated using statistical analysis, and the true scores of a report is added to the coefficient of evaluation tendencies of an evaluator who has not read the report, the estimated value that could be obtained if the evaluator had read and evaluated the report can be calculated. The mean value that is assumed to be obtained if all the evaluators read and evaluated the report can be then calculated. Since the total evaluation tendencies of each evaluator is adjusted to be nil, the mean value of all the evaluators' ratings correspond to the true score.

One external evaluator and two internal evaluators read all 48 reports and the rest of the evaluators read about 10 reports from fiscal 2002 and 2003. This means that each report was read by at least two external evaluators and three internal evaluators. Then, analysis was performed using the above-mentioned statistical analysis method*. The same analysis method was used for the 10 terminal evaluations from fiscal 2001, whose reports were read by one external evaluator and one internal evaluator, respectively. The internal evaluators were not assigned to the secondary evaluation of the projects of the same department or ones that they were previously in charge of.

Table 4-2 Secondary Evaluation Check Sheet

Terminal Evaluation on (Project Title)
Rating criteria
1) The terminal evaluation report is reviewed in light of aspects considered essential as a good evaluation, and each aspect is rated using **"A" as Good, "B" as Average, and "C" as Poor** in green cells. [sub-criteria]
2) The above aspects are classified into several key components. Each component is also scored based on the 1 to 10 scale in yellow cells. [evaluation criteria] **10 and 9: Excellent** **8 and 7: Good** **6 and 5: Average** **4 and 3: Poor** **2 and 1: Very Poor**

*See Technical Notes at the end of chapter 2 (*1 Estimation of evaluation scores).

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 10

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		Rating
Viewpoint	The evaluation team members are selected on an impartial basis.	
3. Evaluation Team Composition—Specialty		Rating
Viewpoint	The evaluation team members are selected with balanced specialty.	
4. Level of Counterpart Participation		Rating
Viewpoint	The counterparts understand evaluation process, and share responsibilities for evaluation activities with JICA.	
Comment		Overall 10

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. Data Collection Methods (*1)		Rating
Viewpoint	Several different data collection methods are used to increase accuracy and reliability of the data/information obtained.	
3. Data/Information Sources		Rating
Viewpoint	The sources of the data/information are adequately explained in the evaluation report.	
4. Appropriateness of Data Sources		Rating
Viewpoint	The data/information is obtained from a broad range of stakeholders including the primary beneficiary groups to limit bias of the data collected.	
5. 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 10

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 (*2)		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 Relationship—Logic of Project Design (*3)		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 Relationship—Before and After (*4)		Rating
Viewpoint	The causal relationship is thoroughly examined to verify that effects for the beneficiaries have resulted from the project interventions.	
Comment		Overall 10

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 10

3.3 Evaluation (*5)

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 objective, causal relationship between outputs and project objective, 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 relationship 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. Basis of Evaluation Results		Rating
Viewpoint	The basis and rationales of evaluation results are explained in a convincing manner.	
7. Conclusion		Rating
Viewpoint	The conclusion is drawn based on holistic viewpoints on the basis of the five evaluation criteria.	
Comment		Overall 10

4. Recommendations/Lessons Learned (*6)

1. Relevance of Recommendations		Rating
Viewpoint	The recommendations are based on the information obtained through the process of data analysis and interpretation.	
2. Relevance of Lessons Learned		Rating
Viewpoint	The lessons learned are based on the information obtained through the process of data analysis and interpretation.	
3. Convincing Recommendations		Rating
Viewpoint	The recommendations and their rationales are objective and convincing.	
4. Convincing Lessons Learned		Rating
Viewpoint	The lessons learned and their rationales are objective and convincing.	
5. Sufficiency of Recommendations		Rating
Viewpoint	The recommendations consider all the impeding/promoting factors identified during the evaluation process.	
6. Sufficiency of Lessons Learned		Rating
Viewpoint	The lessons learned consider all the impeding/promoting factors identified during the evaluation process.	
7. Usability of Recommendations		Rating
Viewpoint	The recommendations are practical and useful for feedback and follow-ups, with a specific time frame.	
8. Usability of Lessons Learned		Rating
Viewpoint	The lessons are generalized and conceptualized so that they are widely applicable.	
Comment		Overall 10

5. Reporting

1. Presentation/ Legibility		Rating
Viewpoint	The evaluation report is simple and clear and understandable to readers—in light of the structure, font, terminology, and data presentation.	
2. Clarity		Rating
Viewpoint	Logical structure and major points are clearly described in an easily understandable manner.	
3. Utilization of Tables and Figures		Rating
Viewpoint	Tables and figures are effectively utilized to visually present statistics and analysis results.	
4. 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 10

III. General Criteria for Good Evaluation (*7)

1. Usefulness		Rating
Viewpoint	In light of the effective feedback to the decision-making of the organization, clear and useful evaluation results are obtained.	
2. Impartiality and Independence		Rating
Viewpoint	Evaluation is impartially conducted in a neutral setting.	
3. Credibility		Rating
Viewpoint	In light of the specialties of evaluators, transparency of the evaluation process, and appropriateness of the criterion of judgment, evaluation information are credible.	
4. Participation of the Partner Country		Rating
Viewpoint	Partner countries' stakeholders participate actively in the process of evaluation, and do not just provide information.	
Comment		Overall 10

IV. Evaluation of the Project Based on the Report

1. Relevance		
Comment		Overall 10
2. Effectiveness		
Comment		Overall 10
3. Efficiency		
Comment		Overall 10
4. Impact		
Comment		Overall 10
5. Sustainability		
Comment		Overall 10
6. Overall Evaluation		
Comment		Overall 10

V. Overall Comment

Notes:

*1 Major data collection methods
1. Literature review
2. Direct observation
3. Questionnaire survey
4. Interview survey
5. Focus group discussion

*2 Information to be gained through examination of implementation process
1. Examination of activities
2. Relationship with target group
3. Project management
4. Overall viewpoint

*3 Qualitative approach to analyze causal relationships
1. Construct information on implementation process from inputs through activities to outputs, and from outputs to objectives
2. Attempt to explain the logical relationship between project implementation and effects
3. Analyze the process to transfer and disseminate technologies through activities
4. Clarify the relationship between project implementation and effects by conducting detailed and in-depth survey of a target region or a target group of small size (e.g. case study)

*4 Quantitative approach to analyze causal relationships
1. See changes of a target society/ beneficiary after the project
2. Compare a target society/ beneficiary with another society/ beneficiary without the project
3. Combination of 1 and 2 (experimental design method)
4. Combination of 1 and 2 (quasi- experimental design method)

*5 Refer to Chapter 2, Part 3 of the JICA Guideline for Project Evaluation as for the viewpoints regarding five evaluation criteria.

*6 Definition of Recommendation and Lessons Learned
Recommendation: include specific measures, suggestions, and advice on a target project for JICA or those concerned in the implementation agencies
Lessons Learned: can be learned through the experience of a target project and fed back to on-going similar projects or to project finding and planning process in the future

*7 Refer to Chapter 1, Part 1 of the JICA Guideline for Project Evaluation as for the definitions of Criteria for Good Evaluation

Remark:
Revised JICA Guideline for Project Evaluation: Practical Methods for Project Evaluation (2004) compiled by the Office of Evaluation, Planning and Coordination Department and published by Japan International Cooperation Publishing Co., Ltd.
This guideline is available on the Evaluation page of JICA website (www.jica.go.jp/english/evaluation/index.html).

Evaluation Results

2-1 Examination of the Quality of Primary Evaluation

(1) Current Conditions of Evaluation Quality and Challenges

1) Overview of Evaluation Results

Average scores for individual evaluation criteria of the 48 terminal evaluations conducted in fiscal 2002 and 2003 are shown in Figure 4-2. All the average scores are over 6.0 points and belong to the level of "average" (≧5 and < 7) in the grading scale. The scores are relatively high for the criteria of "analysis/evaluation (evaluation)" and the general criteria for good evaluation; however, the average scores for "evaluability" and "recommendations/lessons" are rather low.

When looking at the distribution of scores by evaluation criteria, as shown in Figure 4-3, many are distributed between 5.0 and 8.0 as a whole. However, most of the scores for "eval-

Figure 4-2 Average Scores by Evaluation Criteria

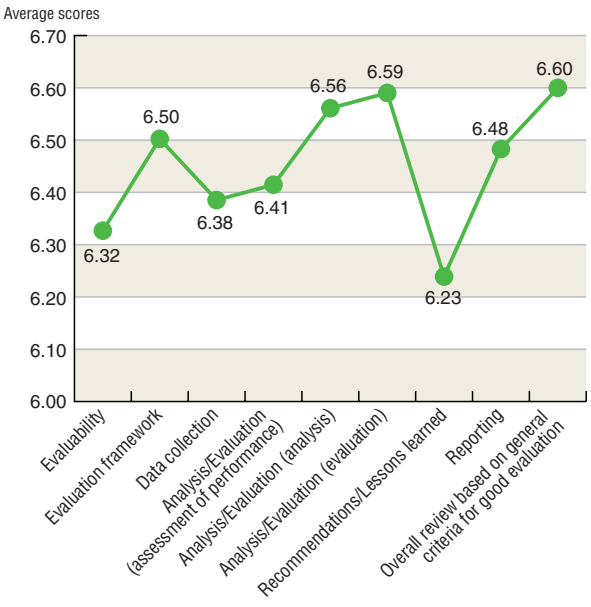
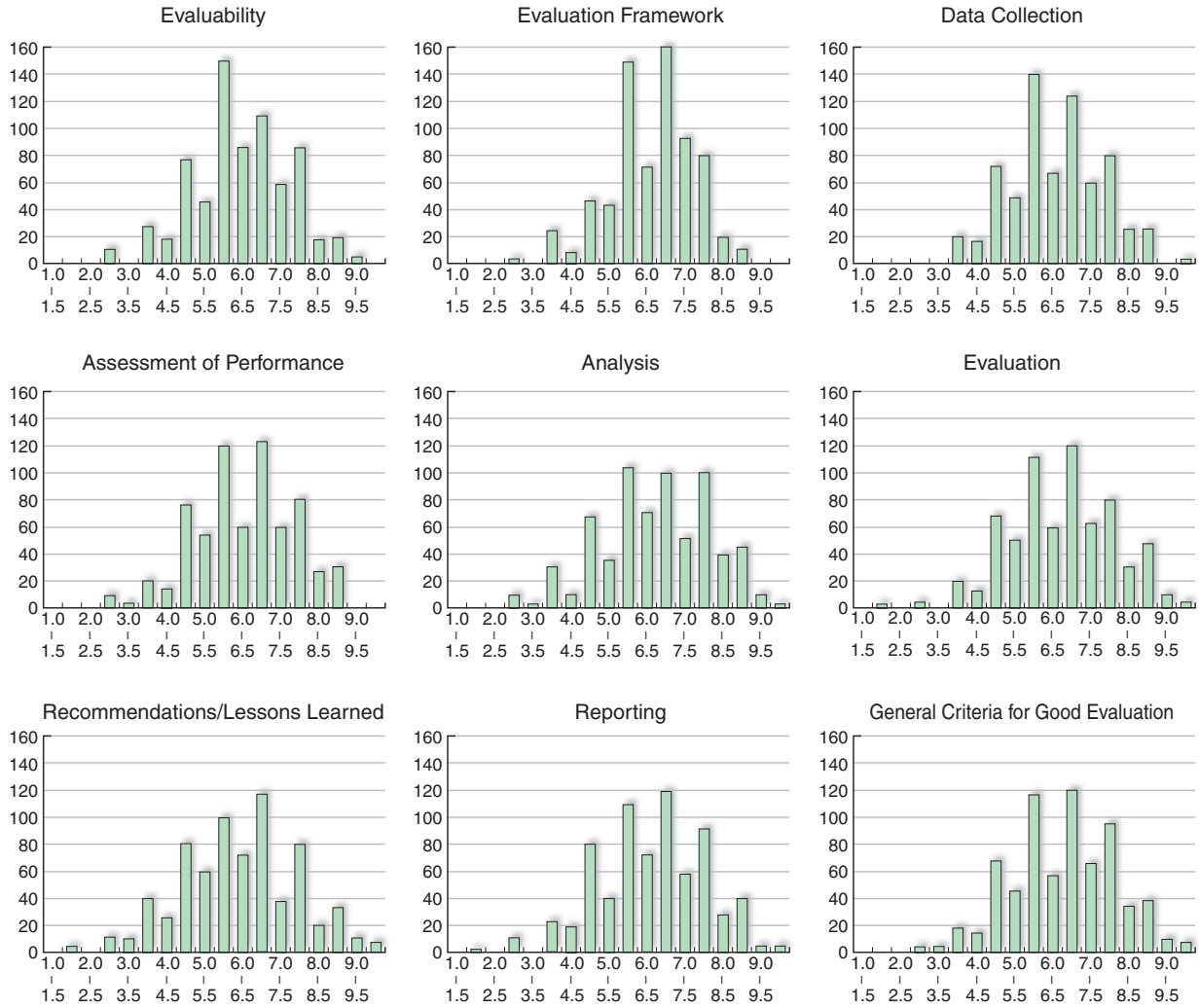


Figure 4-3 Distribution of Scores by Evaluation Criteria



Note: The vertical axis indicates the number of projects subject to the secondary evaluation, where the population parameter is 672, which is obtained when all 14 evaluators read all 48 reports.

uation framework" are between 6 and 7 points with little variance, whereas looking at the scores for each criterion of "analysis/evaluation" and "recommendations/lessons learned," some of the ratings are over 8 points and others below 5 points, thus demonstrating a variance in the quality of primary evaluation. In particular, relatively many projects were graded less than 5 in "recommendations/lessons learned." Most scores for "evaluability" fall near the average score without much variance.

Based on the above results, the following can be said about the quality of the primary evaluation of the target projects. Many scores are clustered around "average" (≧ 5 and < 7) and "slightly good" (≧ 7 and < 9), thus meeting the average level and a certain quality on average. However, there are variances in the quality of primary evaluation for some criteria, represented by "analysis/evaluation" and "recommendations/lessons learned"; in particular, the quality of "recommendations/lessons learned" is relatively low.

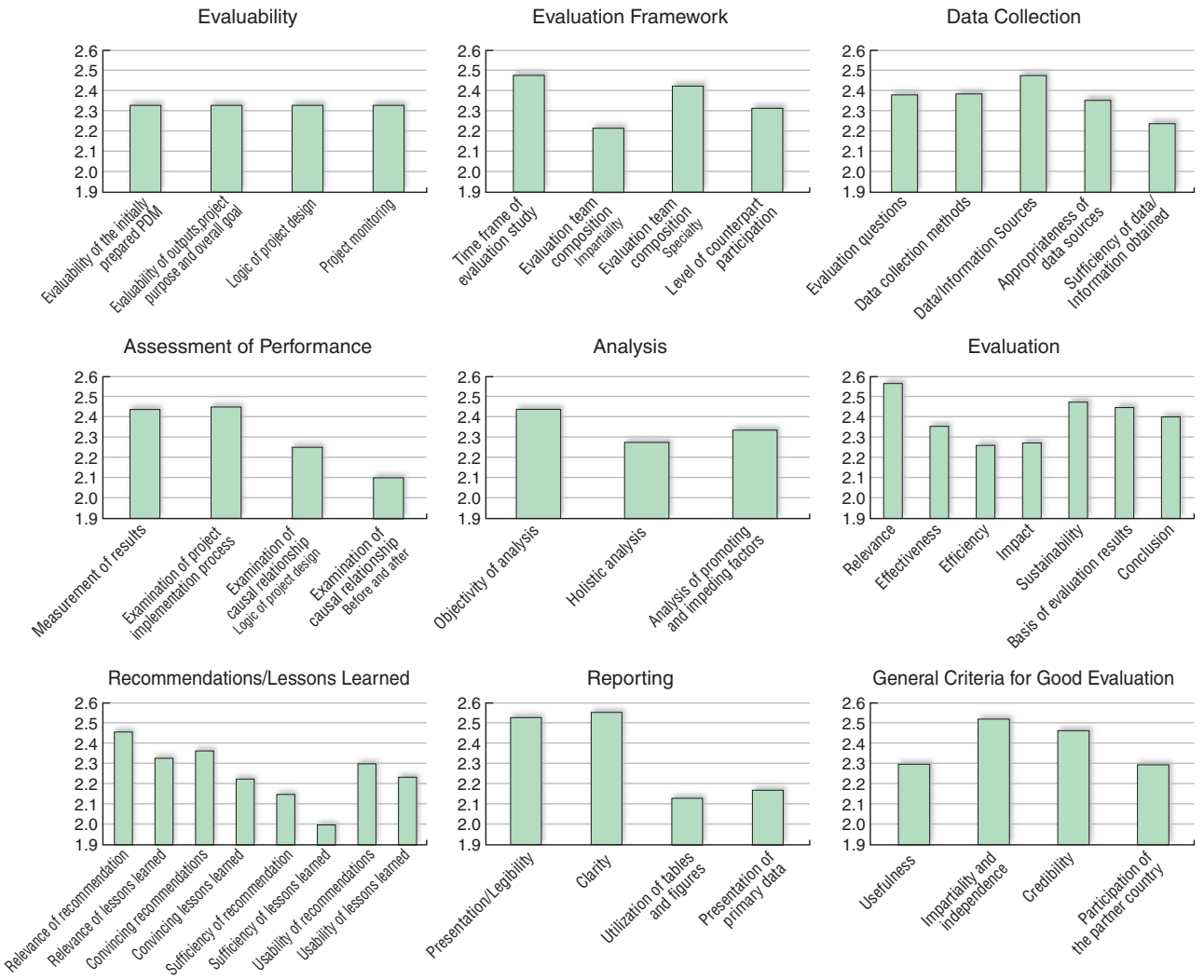
2) Evaluation Results by Criteria and Issues for the Improvement of Quality

In the secondary evaluation, the sub-criteria of each evaluation criterion were rated, and qualitative evaluation information was collected in the form of comments that were written in the additional box by the evaluators. We will take a look at the evaluation results of each evaluation criterion and the current conditions of the quality of evaluation by means of ratings of the sub-criteria and specific issues based on comments from the members. Figure 4-4 shows the evaluation results (average scores) of the sub-criteria of each evaluation criterion. Sub-criteria are evaluated on a three-grade scale—A, B and C. For the convenience of analysis, each grade is converted numerically into 3, 2, and 1, respectively.

a. Evaluability

The scores of the sub-criteria in the category of evaluability are in the range of 2.2- and 2.3-points, and they are not particularly high or low compared with the sub-criteria of the other evaluation criteria. However, the scores for "evaluabili-

Figure 4-4 Evaluation Results of Sub-criteria of Each Evaluation Criterion



Note: The vertical axes indicate rating scores.

ty of the initially prepared project design matrix (PDM)", and "project monitoring" are relatively low. The first secondary evaluation revealed that there were cases where PDM had to be changed drastically at the evaluation stage in order to close the gap between the PDM and the actual project, which was generated due to the lack of appropriate purposes and indicators at the planning stage. In response to this assertion, a criterion for evaluability of the initially prepared PDM was added. Evaluations are made to examine the degree of achievement by comparing it with the plan, and this criterion asks whether the plan has gone through many changes at the evaluation stage. The criterion of "project monitoring" asks whether information and data necessary for evaluation were accumulated through periodical monitoring, and this was also pointed out by the first secondary evaluation.

This secondary evaluation found many cases where PDM was reviewed whenever necessary during the implementation stage, and there were few major changes in PDM at the time of the terminal evaluation. However in some cases, the overall goal was lowered because of a gap between the overall goal and the project purpose. Some assert that the concept of changing the overall goal itself is the problem, because overall goals are, in nature, challenges to be overcome eventually by the project. Moreover, many pointed out that the setting of overall goals was inadequate (setting overall goals too high or too low). In some cases, even though there were changes in PDM during the implementation process, no information regarding the details of the change and the changing process was described or included. The most common comment from the members had to do with problems with the data necessary for evaluation. Problems were specifically pointed out regarding the insufficient data accumulation of indicators to measure the achievement, the inappropriateness of indicators set originally, and the lack of baseline data.

b. Evaluation Framework

The scores of the sub-criteria in the evaluation framework category are generally high, including "time frame of evaluation study." However, the score for the composition of the study team (impartiality) is somewhat low. This is because most of the members of the study team are involved in the project, including project stakeholders from Japanese partner organizations. The participation of stakeholders who have provided actual technical assistance in the project can be an advantage in terms of specialty. However, when all the evaluators other than evaluation consultants are parties involved in the project, independence may be compromised. Thus, it was pointed out that a structure to include the participation of third parties should be considered.

As far as study team members are concerned, many expressed the opinion that the provision of names and affiliations in the report was not enough, and it was hard to understand how they were involved in a given project although it was assumed from their affiliations that they were concerned parties in one way or another. Many also pointed out that it was not possible to identify the specialties and job descriptions of the members based on their affiliations, and it was hard to determine the appropriateness of the number and composition of the staff. With regard to the participation of concerned parties in the partner country, every primary evaluation is conducted jointly. A joint evaluation committee was set up and both parties signed the joint evaluation results. However, there were comments that it was difficult to judge how much the partner country was involved in the evaluation, specifically whether evaluation was carried out jointly, whether concerned parties in the partner country were simply interviewed by the evaluation team from Japan, and whether they simply participated in the evaluation committee. Many asserted that more information about the evaluation framework should be included to make it explicit as to whether or not it is appropriate.

Many commented that the study period in the schedule seemed short. As a successful example of efficient study, there was one case where a part of the team started the preliminary survey in the project site beforehand, and later they were joined by the rest of the team members.

c. Information Collection

The scores in this evaluation criterion are in the range of 2.3- and 2.4-points, except for "sufficiency of data/information obtained." The highest score is recorded in the sub-criterion of "data/information sources," which asks whether the sources of information are identified adequately, such as the places of visits, interviewees, and sources of documents. Some primary evaluations did not sufficiently specify the sources of documents, but it was evaluated to be appropriate as a whole.

The criterion of "sufficiency of data/information obtained" that scored relatively low asks whether information necessary for evaluation was sufficiently collected. Common comments on information collection include a lack of data to verify the degree of achievement, and inadequate information collection about the effects over development challenges (overall goal) in contrast with information collection biased toward the activities and outputs. In addition, in relation to the sub-criterion of "appropriateness of data sources," many also commented that there was insufficient information gathered from a broad range of sources, particularly from the beneficiaries whose information was important for evaluation. It was also pointed out that the sources of information depended heavily on stakeholders and organizations involved in the project, represented by cases where data were collected from the records of the project alone or interviews were conducted only with the parties involved in the project, such as counterparts.

d. Assessment of Performance

Both of two sub-criteria, "measurement of results" and "examination of project implementation process," in this evaluation criterion were rated in the 2.4-point range. However, low scores were given to the two sub-criteria in relation to causal relationship, one that verifies the framework of the project at the planning stage, and the other that verifies the relationships between the effects and project implementation through comparison of with/without and before/after.

Many commented that there were few evaluations that quantitatively verify the effects of a project by comparing data between before and after the project and between target and non-target areas. Many also pointed out that examination of the effects on the overall goal was insufficient. In this regard, inappropriate setting of overall goals, insufficient collection of data as indicators for the effects on the overall goals, and inadequate logical presentation on the causal relationship between implementation of the project and achievement of overall goals were referred to as problems. Furthermore, in some primary evaluations, the achievement of outputs and purposes were measured using only indicators without sufficient examination of the logical framework and the implementation process, resulting in a failure to analyze promoting and impeding factors. Nonetheless, others found other primary evaluations to be exemplary: for example, quantitative evaluations of effects by comparing the baseline and indicators and observing year-to-year changes in indicators. Another good example was the evaluation that analyzed promoting and impeding factors by combining both quantitative and qualitative information.

e. Analysis

The score of "objectivity of analysis" is in the 2.4-point range, whereas "holistic analysis" scored somewhat low. Many commented that the analysis was superficial; specifically, that the analysis was weak due to a lack of information and data; that the analysis was insufficient as it depends solely upon limited sources of information such as interviews, etc.; that there was little quantitative analysis; and that there was a lack of expertise required for the analysis. One primary evaluation was criticized as being a mere description of phenomena or an outline of facts, instead of an analysis. The sub-criterion of holistic analysis determines whether an analysis was performed from various angles by combining several data analysis methods, however, their scores are low reflecting the above comments.

It is worth mentioning that several primary evaluations gave 8.0 points or more to the criteria of "analysis" in the secondary evaluation. Such primary evaluations with high scores performed balanced analyses, using qualitative and quantitative methods based on various kinds of information.

f. Evaluation

Five sub-criteria out of seven, such as "relevance" and "sustainability," were rated at more than 2.3 points, which is higher than other sub-criteria overall. However, the sub-criteria of "efficiency" and "impact" received somewhat low scores.

"Efficiency" received the lowest score among the DAC Five Evaluation Criteria in the first secondary evaluation as well. The low score was mainly due to the insufficient evaluation from a cost-effectiveness point of view. Though a few primary evaluations attempted to analyze cost aspects, there was no case that succeeded in performing a highly convincing cost-effectiveness analysis.

As far as "impact" is concerned, many pointed out that the intended impacts for overcoming some development challenges (overall goals) were not amply evaluated. In this regard, the problem in appropriateness of the set overall goal, the lack of data used as indicators, and the logical issues of causal relationship were referred to. Some commented that there was confusion regarding the definition of impacts; for example, only the unexpected impacts were evaluated in some primary evaluations without evaluating the intended impacts, namely the effects on overall goals.

g. Recommendations/Lessons Learned

The sub-criteria of "relevance" of recommendations/lessons learned, which examine whether recommendations and lessons were extracted through the process of data analysis and interpretation, were rated higher than 2.3 points. And the sub-criterion of "convincing recommendations" were also rated higher than 2.3 points. But the scores of the sub-criteria in this evaluation criterion are generally low compared to those in other criteria. In particular, both sub-criteria concerning "sufficiency" scored the lowest among all the sub-criteria in this check sheet. These sub-criteria determine whether a set of information, including promoting and impeding factors obtained through the process of evaluation, was fully reflected in the recommendations and lessons.

Many comments indicate that the recommendations and lessons are superficial. Particularly with regard to lessons, although many issues on planning and implementation that could provide insights into lessons for the future are revealed in the process of data analysis and interpretation, they are not fully incorporated into the lessons, and those lessons are not systematically drawn out. As far as "usability" is concerned, some primary evaluations presented recommendations according to tasks to be undertaken over the short term or the long term, as well as tasks on the Japanese side or of the partner country. On the other hand, many commented that recommendations and lessons needed to be improved since some recommendations were unclear about by whom and when they would be undertaken, and also since descriptions of some

recommendations and lessons were too general and did not carry specific information.

h. Reporting

The scores for sub-criteria with regard to overall writing style, including report structure, wording, and clarity of logic, fall into the 2.5-point range. However, "utilization of tables and figures" and "presentation of primary data" were rated low, in the 2.1-point range. Some reports incorporated basic statistical data in the main text using easy-to-understand tables and figures, and some reports compiled the results of the questionnaires and interviews into the main text and were attached as references at the end of the publication. On the other hand, there were many reports in which statistical data were listed only in the annex, and although many documents were attached, they did not contain important primary information, such as results of questionnaires, etc. In some reports, large parts were devoted to supplementary materials, containing only a few pages of the main text.

i. General Criteria for Good Evaluation

The scores of these sub-criteria are generally higher than those of other evaluation criteria; those regarding "impartiality and independence" and "credibility" are particularly high. Somewhat lower scores are recorded for "usefulness" and "participation of the partner country," in which problems regarding the participation of the concerned parties of the partner country, information collection from the beneficiaries, analysis, and recommendations/lessons were pointed out as mentioned in other evaluation sub-criteria.

This secondary evaluation examined the relationships between evaluation criteria using statistical analysis*. It was found from the analysis that there is a positive correlation among all the evaluation criteria; particularly strong correlations were observed between "data collection" and each criterion regarding analysis/evaluation ("assessment of performance," "analysis" and "evaluation"); among those three criteria regarding analysis/evaluation; between "recommendations/lessons" and "assessment of performance"; and also between "recommendations/lessons" and "analysis." These results indicate that the secondary evaluation concluded that adequate data collection is vital to high quality analysis/evaluation, and proper assessment of performance as well as analysis are essential for drawing high quality recommendations/lessons.

In the same way, the rating of "general criteria for good evaluation" exhibits strong correlations with all the criteria of "evaluation framework," "data collection," "analysis/evaluation," "recommendation/lessons learned," and "reporting"; a particularly strong relationship is found with "analysis/evaluation" and "recommendations/lessons learned." In other words, good evaluation requires quality in analysis/evaluation and recommendations/lessons.

This analysis also shows that there are positive correlations between "evaluability" and all the key evaluation criteria. Although the correlations with "evaluation framework" and "reporting" are not so strong, it is strong with "data collection" and "evaluation." According to the results of the secondary evaluation, it can be stated that low evaluability has negative impacts on proper data collection for evaluation and on evaluation itself. However, the correlation between "evaluability" and "general criteria for good evaluation" is not as strong as the correlation between the other key evaluation criteria and "general criteria for good evaluation." Therefore, when evaluability is low, it creates difficulties in collecting information, but it does not necessarily reduce the quality of the evaluation. That may be because there are good evaluations that have properly analyzed the issues, including the low evaluability, as seen in some of the target primary evaluations.

3) Examples of Good Quality Evaluation Reports

The JICA Guideline for Project Evaluation explains in detail important points to be considered for appropriate evaluations with regard to key criteria such as evaluation framework, data collection, assessment of performance, analysis, evaluation, recommendations/lessons learned, and reporting. However, general descriptions do not necessarily help with concrete understanding. If high quality evaluation reports are objectively presented using the results of secondary evaluation, these reports can serve as a role model. And if this procedure is repeated, various models will be put together for different sectors and issues. It is then expected that quality will be solidly maintained by actually conducting evaluation studies and compiling reports while referring to the methods and contents in those models.

Under this concept, among the 48 terminal evaluations subject to secondary evaluation in fiscal 2004, the following four terminal evaluation reports were selected to serve as models for others. Each report scored an average of more than 7.5 points in criteria such as key evaluation criteria and general criteria for good evaluation, which were associated with the quality of evaluation itself, and scored no less than 6.0 points in all other criteria, resulting in well-balanced, high quality evaluation reports on the whole. Many scored more than 8 points in understanding of "assessment of performance," "analysis/evaluation," and "recommendations/lessons learned." They commonly came out with a convincing evaluation and through thorough performance assessment and analysis based on both qualitative and quantitative data, and drew

*See Technical Notes at the end of chapter 2 (*2 Correlation among evaluation criteria).

specific and highly useful recommendations and lessons.

One promoting factor may be that the primary evaluators who performed evaluation studies had superior knowledge and understanding of evaluation and specialty areas. They were also well versed in evaluation methods. At the same time, each evaluator was able to obtain adequate data necessary for assessment of performance and analysis, which might have been due to appropriate monitoring and the accumulation of necessary data during the implementation stage of the project, which could have been a promoting factor to the high quality evaluation.

Although some difficulty with the project itself was found in the Groundwater Development and Water Supply Training Project in Ethiopia, its primary evaluation earned the highest scores in many criteria. The difficulty was precisely and frankly addressed based on very convincing grounds after the performance assessment of the project and a thorough analysis of the promoting and impeding factors. That is why many secondary evaluators acknowledged it as a high quality evaluation.

Terminal Evaluation Reports Rated as High Quality Evaluations by the Secondary Evaluation

Ethiopia: The Groundwater Development and Water Supply Training Project

Jordan: Information Technology Upgrading Project

Jordan: The Project for Family Planning and Gender in Development

Kenya: Strengthening of Mathematics and Science in Secondary Education

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

We have thus far examined the quality of evaluation by means of evaluation criteria targeting 48 terminal evaluations in fiscal 2002 and 2003. Adding terminal evaluations of fiscal 2001 to these 48 cases, the average scores are computed by the fiscal years and by the criteria shown in Table 4-3 and Figure 4-5 respectively. In each fiscal year, as explained in the discussions of evaluation method, the scores are estimated on the assumption that the target primary evaluation reports were read and evaluated by all the secondary evaluators, and the average scores are calculated accordingly. Since the reports in fiscal 2001 are used for additional analysis to compare with the results in fiscal 2002 and 2003 and the number of the secondary evaluators is different, the scores are showed in a different way with dotted lines.

The table and figure show an increase in scores for "evaluation framework," "data collection," and "analysis/evaluation" (including "assessment of performance," "analysis," and "evaluation") by the 0.1- and 0.2-point range from fiscal 2002 to 2003, although there are some fluctuations between the fiscal years. In particular, the score of analysis increased by the 0.2-point range for three years in a row. However, the scores of "recommendations/lessons learned," "reporting," and "general criteria for good evaluation" basically remain unchanged. The score of "evaluability" in fiscal 2003 dropped from what it was in fiscal 2002.

The above shows the results of the primary evaluation of targeted projects alone, not the results that show changes in the overall terminal evaluations in a given fiscal year. The targeted projects in fiscal 2003 and 2001 are 10 primary evaluations for each year, and it is hard to tell whether the same results would have been obtained from the secondary evaluation targeting all the terminal evaluations. Thus, in order to identify whether there were year-to-year changes in quality, the secondary evaluation was carried out using statistical analysis if all the terminal evaluation results in a given year had been evaluated*. Table 4-4 shows the summary of the analysis results. The criteria between which differences in the average scores were inferred are indicated with circles. When

Figure 4-5 Year-to-Year Changes in the Quality of Evaluation

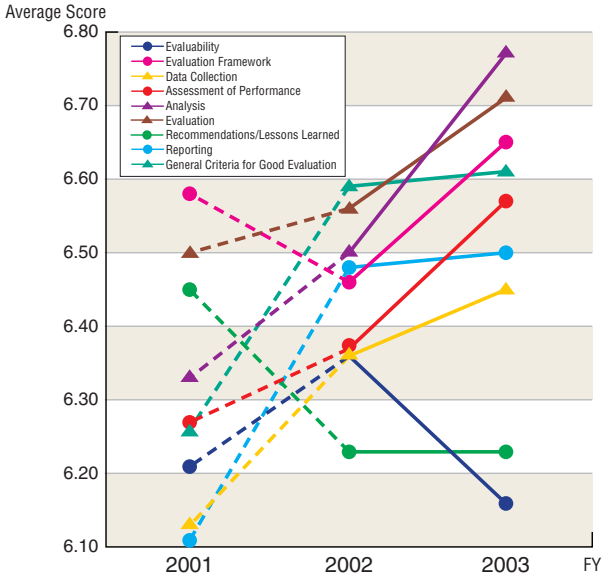


Table 4-3 Score Changes by Fiscal Year and Criteria

FY	Evaluability	Evaluation framework	Data collection	Assessment of Performance	Analysis	Evaluation	Recommendations/ Lessons learned	Reporting	General Criteria for Good Evaluation
2001	6.21	6.58	6.13	6.27	6.33	6.50	6.45	6.11	6.26
2002	6.36	6.46	6.36	6.37	6.50	6.56	6.23	6.48	6.59
2003	6.16	6.65	6.45	6.57	6.77	6.71	6.23	6.50	6.61

*See Technical Notes at the end of chapter 2 (*3 Examination of the differences in quality between fiscal years).

there are significant differences, the fiscal year with the higher score is presented in brackets.

This analysis result indicates that the mean scores of the terminal evaluation in the criterion of "evaluability" show no significant differences from fiscal 2001 to 2002, from fiscal 2002 to 2003, and from fiscal 2001 to 2003.

The mean scores for "data collection," "reporting," and "general criteria for good evaluation" in fiscal 2002 are higher than those in fiscal 2001; the scores for "evaluation framework" and "analysis" in fiscal 2003 are higher than those in fiscal 2001: and those for "data collection," "assessment of performance," "analysis," "reporting," and "general criteria for good evaluation" in fiscal 2003 are higher than those in fiscal 2001. There are no particular differences in the mean scores for "evaluation" and "recommendations/lessons learned" between the fiscal years.

It is fair to conclude that the result shows a year-to-year increase in the quality of evaluations in all criteria except for "recommendations/lessons learned" and "evaluability."

(3) Difference in Perspectives between External Evaluation and Internal Evaluation

1) General Tendency

This secondary evaluation examined the differences in evaluation results between external evaluators and internal evaluators. Table 4-5 and Figure 4-6 show the mean scores of the secondary evaluations of the 48 terminal evaluations for each evaluation criterion conducted by external evaluators and internal evaluators in fiscal 2002 and 2003.

Evaluation tendencies are similar between external evaluators and internal evaluators, and the evaluation scores of external evaluators tend to be higher. The score differences for "evaluation framework" and "general criteria for good evaluation" between both groups of evaluators are smaller than those for other criteria.

Differences between external evaluations and internal evaluations are examined for the secondary evaluations on all the terminal examinations in the target years in the same manner as the year-to-year changes in the quality using statistical analysis, and the results are listed in Table 4-6*. The criteria with differences are indicated with circles and the group of evaluators who gave higher scores is placed in brackets. Consequently, it is found that the evaluation scores of the external evaluators are higher than those of the internal evaluators for all the criteria except for "evaluation framework" and "general criteria for good evaluation."

2) Difference in the Perspectives of Evaluation

In order to further analyze the difference in perspectives between external evaluators and internal evaluators, the mean scores of sub-criteria of each evaluation criterion are compared and shown in Figure 4-7.

As the figure shows, the sub-criteria scores of the external evaluators are generally higher. However, there are some sub-

Table 4-4 Examination Results of Year-to-Year Changes in Quality

Questions	2001/2002	2002/2003	2001/2003
I. Preconditions for conducting appropriate evaluation (evaluability of target projects)			
Evaluability	–	–	–
II. Key evaluation criteria			
Evaluation Framework	–	○ (03)	–
Data Collection	○ (02)	–	○ (03)
Analysis/Evaluation (Assessment of Performance)	–	–	○ (03)
Analysis/Evaluation (Analysis)	–	○ (03)	○ (03)
Analysis/Evaluation (Evaluation)	–	–	–
Recommendations/Lessons Learned	–	–	–
Reporting	○ (02)	–	○ (03)
III. General Criteria for Good Evaluation			
General Evaluation	○ (02)	–	○ (03)

○indicates criteria between which differences in the average ratings were inferred, whereas – indicates criteria between which no differences were inferred.
(02) shows the group with higher mean scores of fiscal 2002, and (03) shows the group with higher mean score of fiscal 2003.

Figure 4-6 Differences in Evaluation Scores between External Evaluation and Internal Evaluation

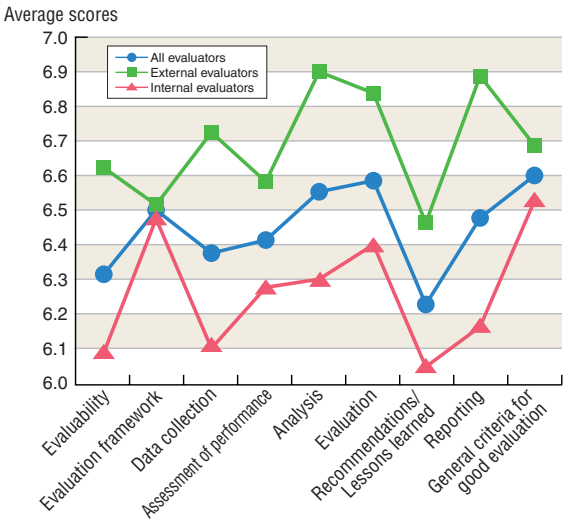


Table 4-5 Differences in Evaluation Scores between External Evaluation and Internal Evaluation

	Evaluability	Evaluation framework	Data collection	Assessment of performance	Analysis	Evaluation	Recommendations/ lessons learned	Reporting	General criteria for good evaluation
All evaluators	6.32	6.50	6.38	6.41	6.56	6.59	6.23	6.48	6.60
External evaluators	6.62	6.50	6.73	6.58	6.89	6.84	6.46	6.89	6.68
Internal evaluators	6.09	6.50	6.12	6.29	6.31	6.41	6.06	6.18	6.53

*See Technical Notes at the end of chapter 2 (*4 Examination of the differences in evaluation results between the external evaluators and internal evaluators).

Table 4-6 Differences in Evaluation Scores between External Evaluation and Internal Evaluation

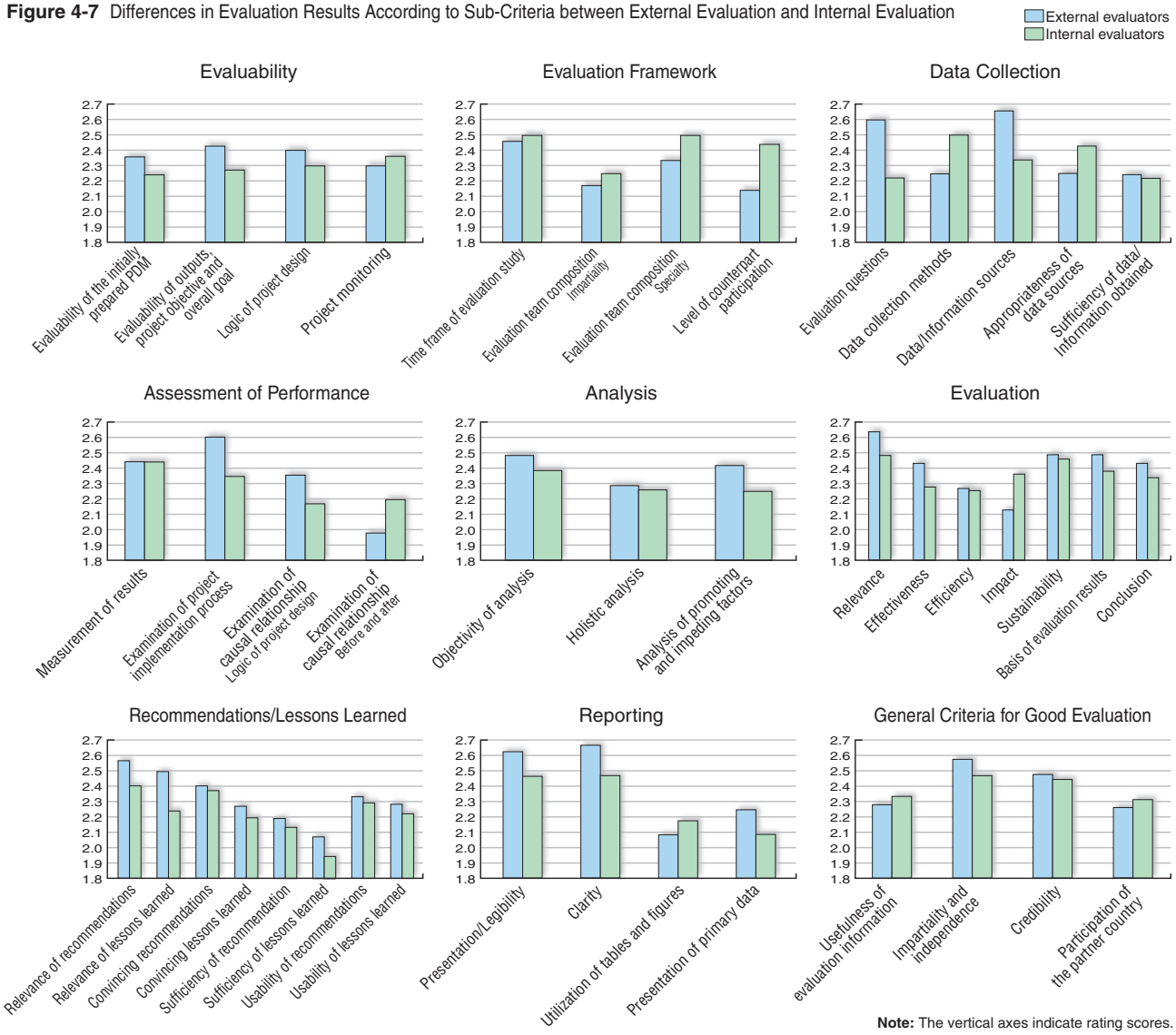
Questions	Existence of Differences in the Mean Scores of the Population
I. Preconditions for conducting apprpriate evaluation (evaluabilty of target projects)	
Evaluability	○ (External)
II. Key evaluation criteria	
Evaluation Framework	–
Data Collection	○ (External)
Analysis/Evaluation (Assessment of Performance)	○ (External)
Analysis/Evaluation (Analysis)	○ (External)
Analysis/Evaluation (Evaluation)	○ (External)
Recommendations/Lessons Learned	○ (External)
Reporting	○ (External)
III. General Criteria for Good Evaluation	
General Evaluation	–

○ indicates criteria between which differences in the mean scores were inferred, whereas – indicates between which no differences were inferred.
(External) indicates the question to which external evaluators have given higher mean scores.

criteria carried out by external evaluation that show lower rating scores, among which all the sub-criteria of "evaluation framework" were given lower scores by external evaluators, and the difference in scores for "participation of the partner country" is particularly large between the external and internal evaluators. Other sub-criteria for which the external evaluators gave lower scores include "data collection methods" and "appropriateness of data sources" in the evaluation criterion of "data collection," "measurement of results" and "examination of causal relationships between effects and project implementation though comparison of before/after and with/without the project" in the criterion of "assessment of performance," "impact" in the criterion of "evaluation," "utilization of tables and figures" in the criterion of "reporting," and "usefulness" and "participation of the partner country" in the "general criteria for good evaluation."

As for "appropriateness of data sources" in the "data collection" criterion, as mentioned already, it was pointed out that there is a lack of information from beneficiaries, which is

Figure 4-7 Differences in Evaluation Results According to Sub-Criteria between External Evaluation and Internal Evaluation



Note: The vertical axes indicate rating scores.

the same perspective indicated in "evaluation framework" or "participation of the partner country" in "general criteria for good evaluation." The above two sub-criteria in "assessment of performance" and the sub-criterion "impact" determine whether effects were verified appropriately. Consequently, it is fair to state that the criteria that received lower rating scores from external evaluators than those received from internal evaluators have something to do with the perspectives of whether the concerned parties in the partner country participated, whether the viewpoints of beneficiaries were reflected in the evaluation, and whether effects of cooperation, particularly on challenges to be overcome (overall goal), were verified based on appropriate data. One of the reasons external evaluators gave lower evaluation scores for data collection methods and the utilization of tables and figures than did internal evaluators may be due to the fact that many of the external evaluators are academics well versed in research and study methodology and thesis writing. Furthermore, internal evaluators gave higher scores to all of the sub-criteria in the criterion of "evaluation framework" than did external evaluators. This may be because study schedules and the composition of study teams are to some extent fixed within JICA, and the internal evaluators may have taken it for granted that there would be little room for changing those schedules and team compositions.

3) Consideration of Differences in Evaluation Tendencies

Several analyses were performed to look into differences in evaluation tendencies for external evaluators and internal evaluators, respectively, using the secondary evaluation results.

First, evaluation results using the key evaluation criteria show that external evaluators tend to give higher evaluation scores. However, this is proved based on a comparison of the mean scores of all the evaluators. In order to determine the tendency of individual evaluators, the scores by criteria of each evaluator are arranged from highest to lowest in Table 4-7. The tendency varies by individual, but it is found that internal evaluators tend to give lower scores in general.

It is generally believed that internal evaluators tend to be more generous in rating; however, the scores of the secondary evaluation revealed that internal evaluators are more severe in rating as a whole. In order to analyze the factors affecting this tendency, an examination was conducted to find how much knowledge each evaluator has about target regions and issues. That is because knowledge about target regions and issues may affect an evaluation tendency.

Accordingly, the evaluators who performed actual secondary evaluations were asked to rate their familiarity with the target regions and issues for each primary evaluation, using a four-grade scoring system: (four points=very familiar, three points=somewhat familiar, two points=have general knowledge, one point=no knowledge at all). Then, mean values for external evaluators and internal evaluators were calculated. The degree of familiarity with the regions of external evaluators, expressed in points, was 2.53 and that of familiarity with the issues was 2.27. The scores for internal evaluators were 1.97 and 1.79, respectively. The external experts' familiarity with regions and issues shows a greater standard deviation and is more widely distributed.

Table 4-7 Evaluation Tendencies of External Evaluators and Internal Evaluators

Ranking	1	2	3	4	5	6	7	8	9	10	11	12	13	14
I. Evaluability														
Evaluability	Internal-E	External-C	External-D	External-B	Internal-H	External-E	Internal-F	External-A	Internal-C	Internal-A	External-F	Internal-D	Internal-G	Internal-B
II. Key Evaluation Criteria														
Evaluation Framework	Internal-H	Internal-G	Internal-E	External-B	External-E	External-F	External-C	Internal-F	Internal-C	External-D	Internal-A	Internal-B	External-A	Internal-D
Data Collection	External-D	Internal-E	External-B	External-C	Internal-H	External-E	Internal-C	Internal-G	External-F	Internal-A	Internal-F	External-A	Internal-B	Internal-D
Analysis/Evaluation (Assessment of Performance)	External-B	External-E	Internal-H	Internal-E	External-C	Internal-G	External-D	Internal-C	Internal-F	External-A	Internal-A	Internal-D	Internal-B	External-F
Analysis/Evaluation (Analysis)	Internal-H	External-B	External-D	External-E	External-C	Internal-E	Internal-G	Internal-C	External-A	Internal-D	Internal-A	External-F	Internal-F	Internal-B
Analysis/Evaluation (Evaluation)	Internal-H	Internal-E	External-B	External-C	External-D	External-E	External-A	Internal-C	Internal-G	Internal-F	Internal-A	External-F	Internal-D	Internal-B
Recommendations/ Lessons Learned	External-E	External-C	Internal-G	Internal-E	External-D	Internal-H	External-B	Internal-C	Internal-F	Internal-A	Internal-D	External-F	External-A	Internal-B
Reporting	External-B	External-A	External-D	Internal-H	Internal-E	External-E	External-C	Internal-G	Internal-C	Internal-A	Internal-F	Internal-B	External-F	Internal-D
III. General Criteria for Good Evaluation														
General Evaluation	Internal-H	External-E	External-B	Internal-G	External-D	Internal-E	Internal-C	External-C	External-F	Internal-F	Internal-A	Internal-D	External-A	Internal-B
IV. Evaluation of the Project Based on the Report														
Relevance	Internal-H	Internal-E	External-B	External-C	External-E	External-D	External-A	Internal-G	Internal-C	External-F	Internal-A	Internal-F	Internal-B	Internal-D
Effectiveness	Internal-H	External-B	External-D	External-E	Internal-E	External-C	External-A	Internal-C	External-F	Internal-G	Internal-A	Internal-F	Internal-D	Internal-B
Efficiency	Internal-H	External-E	External-A	Internal-E	External-C	External-B	External-D	Internal-C	Internal-G	Internal-A	Internal-F	External-F	Internal-D	Internal-B
Impact	External-D	External-B	Internal-H	External-E	Internal-E	External-A	Internal-G	External-C	Internal-C	Internal-A	Internal-D	Internal-F	Internal-B	External-F
Sustainability	External-D	Internal-H	External-C	External-B	External-A	Internal-E	External-E	Internal-C	Internal-D	External-F	Internal-G	Internal-A	Internal-F	Internal-B
Overall Evaluation	Internal-H	External-D	External-B	External-E	Internal-E	External-C	External-A	Internal-C	Internal-G	Internal-F	Internal-A	External-F	Internal-D	Internal-B

Table 4-8 The Degree of Familiarity with Regions/Issues and Influence on Rating Scores

Type of Analysis	Degree of Familiarity with Regions		Degree of Familiarity with Issues	
	External Evaluator	Internal Evaluator	External Evaluator	Internal Evaluator
I. Evaluability				
Evaluability	–	–	–	–
II. Key Evaluation Criteria				
Evaluation Framework	–	○ (Strict)	–	○ (Strict)
Data Collection	–	–	–	–
Analysis/Evaluation (Assessment of Performance)	–	○ (Strict)	–	–
Analysis/Evaluation (Analysis)	–	–	○ (Strict)	–
Analysis/Evaluation (Evaluation)	–	○ (Strict)	–	–
Recommendations/Lessons Learned	–	○ (Strict)	–	–
Reporting	–	–	○ (Strict)	–
III. General Criteria for Good Evaluation				
General Evaluation	–	○ (Strict)	–	–

○ indicates the presence of influences on the rating score, and – indicates no influence. (Strict) indicates a tendency toward strict evaluation grades.

Next, the degree to which familiarity influenced the rating scores of evaluation criteria was examined using statistical analysis*. The results of the analysis are shown in Table 4-8.

The cases in which the degree of familiarity influenced the evaluation scores are differentiated from cases in which the degree of familiarity has not influenced the evaluation scores in each group of evaluators, and the tendency toward higher scores and the tendency toward lower scores are presented in brackets. The table shows that the influence of the degree of familiarity with regions or issues on evaluation scores is different between external evaluators and internal evaluators. In other words, no causal relationships are found between external experts' familiarity with regions and rating scores. On the other hand, the rating scores tend to be lower when the internal experts are more familiar with the target region, thus exhibiting a tendency toward stricter evaluations. As far as the degree of familiarity with issues are concerned, the more external evaluators are familiar with an issue, the stricter the rating scores on "analysis" and "reporting" become, and the more internal evaluators are familiar, the stricter the scores on "evaluation framework" become. Based on the above-mentioned results, one can assume that the internal evaluators gave relatively strict secondary evaluation grades because many internal evaluators have long working experience and are more familiar with the target regions.

The secondary evaluation also analyzed which evaluation criteria or sub-criteria were weighted heavily by each evaluator using statistical analysis in order to observe differences in evaluation tendencies for external and internal evaluators. Evaluators with similar tendencies were divided into groups**. The result is shown in Figure 4-8. Evaluation tendencies are more similar when the lines are connected further at the left side. The evaluators are tentatively divided into two groups based on this analysis result, which is shown in Table 4-9. When the evaluators are classified into two groups, 75% of the internal evaluators are placed in the same group, whereas the percentage is only 67% for external evaluators, suggesting that internal evaluators have similar evaluation tendencies.

Figure 4-8 Grouping of Evaluators by Evaluation Bias

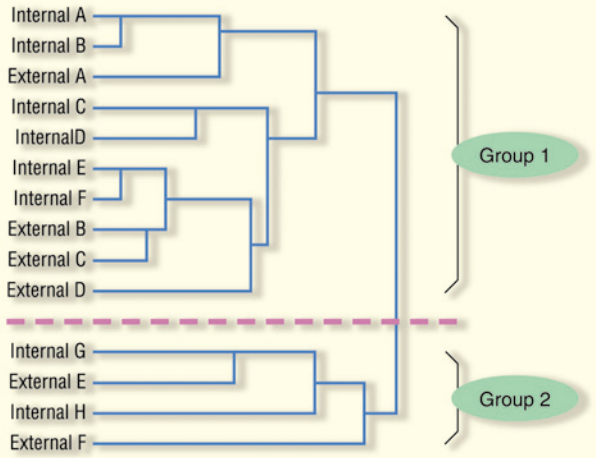


Table 4-9 Classification of Evaluators by Evaluation Bias

Group 1		Group 2	
External Evaluators	**Internal Evaluators**	**External Evaluators**	**Internal Evaluators**
External A, External B, External C, External D (67% of External Evaluators)	Internal A, Internal B, Internal C, Internal D, Internal E, Internal F (75% of Internal Evaluators)	External E, External F (23% of External Evaluators)	Internal G, Internal H (25% of Internal Evaluators)

*See Technical Notes at the end of Chapter 2 (*5 Analysis of the influence of the degree of familiarity on the rating score).
**See Technical Notes at the end of Chapter 2 (*6 Classification of different types of evaluation bias).

It is somewhat obvious that internal evaluators have similar evaluation tendencies and external evaluators have dissimilar tendencies, based on the fact that internal evaluators are engaged in similar work within the same organization, whereas external evaluators have different expertise and backgrounds. Also, having a similar or dissimilar tendency does not impose any problems in themselves. However, based on the above, it is implied that biased evaluation perspectives are possible when evaluations are performed only by internal evaluators who demonstrate high evaluation similarities, and that external evaluators are not free of personal evaluation bias and are not necessarily superior in terms of independence and impartiality. Consequently, it is important to include greater participation from external evaluators in internal evaluations and incorporate various viewpoints into external evaluations so as to obtain evaluation results with greater credibility.

2-2 Secondary Evaluation of Projects Based on the Primary Evaluation

(1) Results of Secondary Evaluation

We have conducted secondary evaluations of 48 projects subject to terminal evaluations in fiscal 2002 and 2003 from the perspectives of the DAC Five Evaluation Criteria and overall evaluation. Since it was a secondary evaluation, it was somewhat limited in terms of available information and differences in the quality of reports. Nonetheless, evaluation results (the mean scores for each criterion) based on the information given in reports are shown in Figure 4-9.

"Overall evaluation" scored 6.34 points, placing it at the "average" level (≧5 and <7) in the rating scale. "Relevance" earned the highest score among the DAC Five Evaluation Criteria, placing it at the level of "slightly good" (≧7 and <9). Four other criteria achieved the "average" level (≧5 and <7); however, the scores on "efficiency" and "sustainability" are slightly low.

Next, we examine the distribution of scores. Figure 4-10 shows the distribution of scores obtained from the calculated mean scores of target projects for each evaluation criterion. Many fall between 5.5 and 8.0 points on the whole. As far as "relevance" is concerned, most of the projects achieved more than 7 points and not a single project got less than 5 points.

Figure 4-9 Mean Scores by Evaluation Criterion

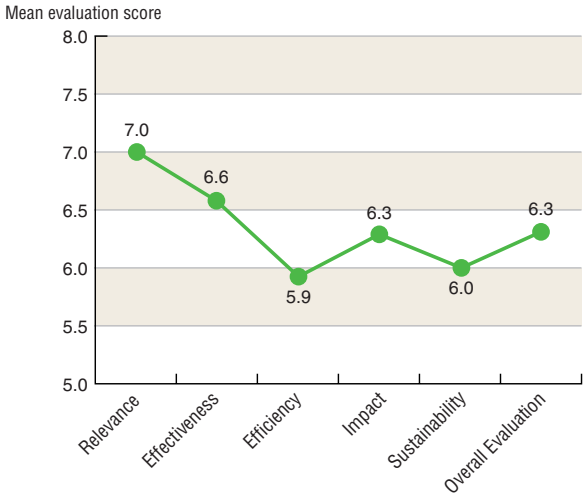
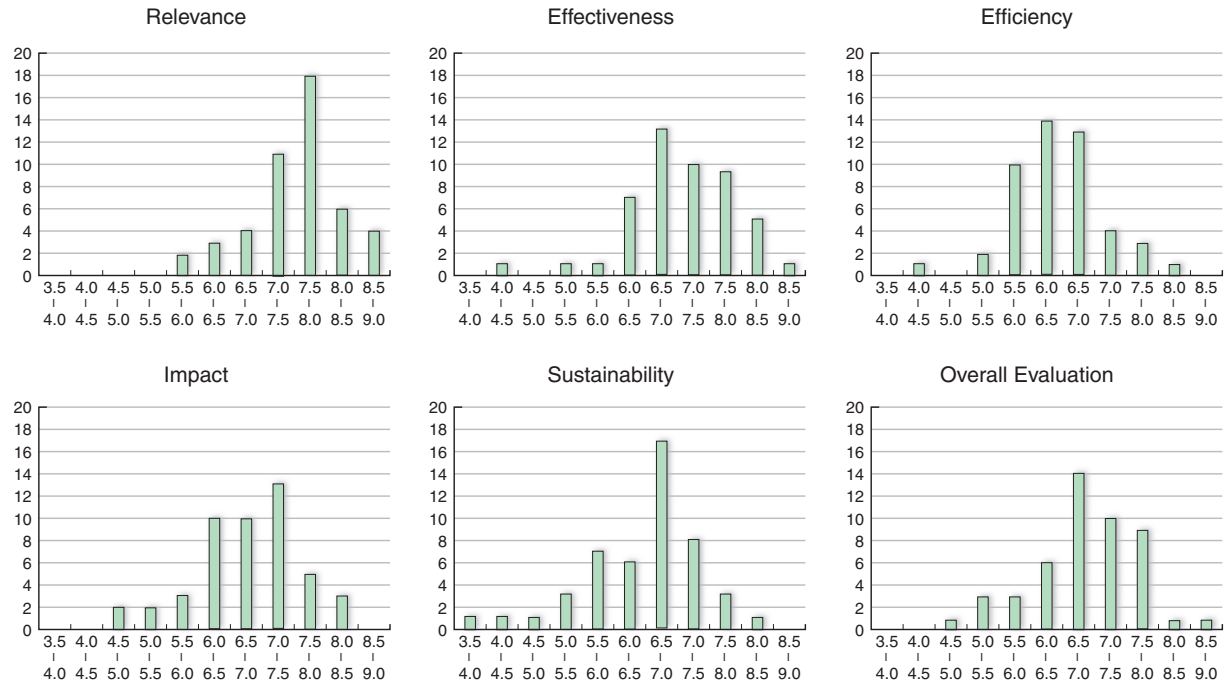


Figure 4-10 Distribution of Scores of Target Projects by Evaluation Criterion



Note: The vertical axes indicate the numbers of projects.

Many scores on "effectiveness" and "impact" are between 6.0 and 7.5 points; some projects scored more than 8 points, whereas others scored less than 5 points. Many scores on "efficiency" fall within the range of 5.5-7.0, showing somewhat low evaluation results. Most scores on "sustainability" are between 6.5 and 7.0; however, they are widely distributed and some received scores in the 3-point range.

(2) Analysis of Evaluation Results

The secondary evaluation revealed that outcomes were different from project to project. In order to analyze whether there are differences in influencing factors on the outcomes of projects depending on the target region or sector, we have classified 48 target projects by region and sector. Figure 4-11 and 4-12 show the mean scores of each evaluation criterion by region and by sector, respectively.

Figure 4-11 Mean Scores of Each Evaluation Criterion by Region

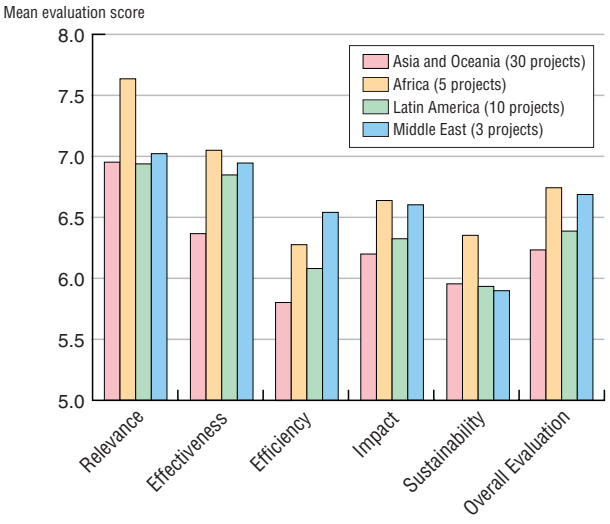
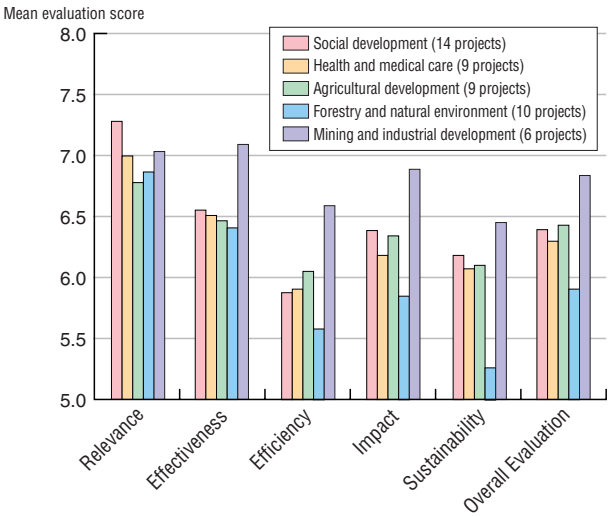


Figure 4-12 Mean Scores of Each Evaluation Criterion by Sector



1) Analysis by Region

The number of target projects is limited in all regions except Asia and Oceania, and it is therefore not appropriate to analyze regional trends based on these results alone. As far as the secondary evaluation in question is concerned, the mean scores for Africa are generally high; the scores for relevance are particularly high. The mean scores for Asia and Oceania are relatively low.

Reasons for high mean scores for Africa may be that there is an extremely successful project among the target projects studied in this secondary evaluation and the other projects also achieved generally good scores with no criteria receiving less than 5 points. The extremely successful project is Strengthening of Mathematics and Science in Secondary Education in Kenya, which achieved the highest scores of the 48 projects in all of the DAC Five Evaluation Criteria and overall evaluation. This is the only project that received more than 8 points for "effectiveness" and "overall evaluation". An example of a project that received "slightly good" ($\geqq 7$ and < 9) for every criterion is the Groundwater Development and Water Supply Training Project in Ethiopia, which was rated at more than 8 points for "relevance" and at more than 7 points for "effectiveness," "efficiency," "impact," and "overall evaluation."

The quality of the primary evaluations for these two projects was also regarded as being high. It is fair to assume that appropriate project implementation, including planning and monitoring, resulted in the high quality of the primary evaluations, which in turn led to good results for the secondary evaluation of the project, as mentioned above.

The evaluation results for Asia and Oceania vary partly due to the fact that there were a large number of projects. The Technical Cooperation Project for Ensuring the Quality of MCH Services through MCF Handbook in Indonesia received scores within the 7-point range on "relevance," "effectiveness," "impact," and "sustainability" for the DAC Five Evaluation Criteria and "overall evaluation." However, the Secondary School Teacher Training Project in Science and Mathematics in Cambodia and the Project for Enhancement of Capabilities in Flood Control and Sabo Engineering of the Development of Public Works and Highways in the Philippines received "poor" ($\geqq 3$ and < 5) ratings for several criteria including "overall evaluation."

The Secondary School Teacher Training Project in Science and Mathematics in Cambodia supported the formulation of mid- and long-term plans for developing and training teachers in secondary science and mathematics, as well as enhancement of the functionality and capacity of the teacher training school, which is meant to improve the quality of basic education in Cambodia. Although the purpose of formulating mid- and long-term plans was mostly achieved, the project activities for the enhancement of functionality and capacity of

the teacher training school had to be changed due to lack of basic capacity of Cambodian teachers and poor training facilities, both of which exceeded initial estimations. The purpose was not fully accomplished during the original period of the project. Consequently, the secondary evaluation gave low scores for "effectiveness," "efficiency," "sustainability," and "overall evaluation," though "relevance" received a high score.

The Project for Enhancement of Capabilities in Flood Control and Sabo Engineering of the Development of Public Works and Highways in the Philippines was implemented to strengthen the system of newly established disaster prevention organizations in the Philippines, where floods and landslides caused by typhoons lead to enormous human suffering and financial problems every year. Focusing on capacity development for the field staff of local offices in particular, this project extended cooperation to improve technical standards, train technicians, and establish a basic information system required for disaster prevention programs. However, inputs were postponed due to chronic budgetary deficits on the Philippine side and an additional survey became necessary due to inadequate or false information related to disaster prevention. The delay of activities prevented the achievement of intended outputs. This led to low scores for "effectiveness," "efficiency," and "sustainability" in the secondary evaluation results. With regard to "relevance," someone commented that the necessity was acknowledged but the priority on the Philippine side was questionable, which is why it received a low score for "overall evaluation."

The cooperation periods for the Secondary School Teacher Training Project in Science and Mathematics in Cambodia and the Project for Enhancement of Capabilities in Flood Control and Sabo Engineering of the Development of Public Works and Highways in the Philippines were extended in response to the terminal evaluation results. Cooperation to achieve the goals is currently underway.

In the Middle East, there are no projects that received less than 5.0 points for each criterion. An example of a project with many "slightly good" ratings (≧ 7 and < 9) is the Information Technology Upgrading Project in Jordan, which received scores in the 7-point range for "relevance," "effectiveness," "efficiency," "impact," and "overall evaluation." The quality of the primary evaluation of this project was also as high as those for the projects in Kenya and Ethiopia mentioned earlier.

The Groundwater Development and Water Supply Training Project in Ethiopia obtained favorable evaluation results in major criteria such as "relevance."

In Latin America, the Project for the Strengthening of Agricultural Technology Development and Transfer in El Salvador received scores in the range of 7- and 8-points for "relevance," "effectiveness," "impact," and "overall evaluation," among which the score for "overall evaluation" was particularly high, following the project in Kenya. In Latin America, the Brazilian Amazon Forest Research Project received "average" (≧ 5 and < 7) grades for "relevance," "effectiveness," and "efficiency," but the grades for "impact" and "sustainability" were slightly low, and the score for "overall evaluation" remained in the 4-point range. This project supported improvement of techniques for forest conservation and restoration of denuded lands with the aim of establishing sustainable utilization of forest resources in the Amazon region, and it transferred techniques to the target research institutions for analyzing characteristics of denuded lands as well as tree types that are important for forest conservation and the rehabilitation of denuded and degraded lands. The intended outputs were accomplished and the capacities of the research institutions increased, thus achieving technical sustainability. However, financial problems due to lack of budget were pointed out, and the score for "sustainability" was low in the secondary evaluation. Furthermore, not a few secondary evaluators mentioned problems with contributions of the project to removing obstacles identified in the first place. While basic research is important, they said, it has not reached the point where technical improvements show a relationship to the utilization of techniques. This led to low scores for "impact" and "overall evaluation."

2) Analysis by Sector

As to analysis by sector, the number of target projects differs from sector to sector, and so it is not possible to analyze trends based on the results of the secondary evaluation alone. However, as far as the 48 target projects are concerned, as Figure 4-12 shows, the mean scores for the projects in the mining and industrial sector are generally high. The mean scores for the forestry and natural environment sector are low as a whole, and sustainability is particularly low.

The reasons for the high average grades in the mining and industry sector is that there are many projects with high evaluation results as a whole. Those projects include the above mentioned Information Technology Upgrading Project in Jordan, the Project on Electrical and Electronics Appliances Testing in the Philippines, the Project on Electric Power Technical Standard Establishment in Laos, and the Project

Quality evaluation was conducted in the Information Technology Upgrading Project in Jordan, which obtained good evaluation results at the same time.

on Upgrading Verification and Inspection Technology in the Area of Mass in Paraguay. These projects received scores in the 7-point range for "overall evaluation," and the scores are in the 6-point range for other criteria.

In contrast, in the field of forestry and natural environment, the score for "overall evaluation" of the Brazilian Amazon Forest Research Project is in the 4-point range, and other projects received "average" (≧5 and <7) grades, which is why the mean scores are relatively low on the whole. The mean scores for "sustainability" were lowered by two projects in Laos with scores in the 4-point range; namely, the Forest Conservation and Afforestation Project in Laos and the Aquaculture Improvement and Extension Projects. Of these two projects, the Forest Conservation and Afforestation Project in Laos received a comment that referred to problems of sustainability in terms of organization and finance, although technological sustainability has been secured. The problems of financial sustainability is also pointed out in Aquaculture Improvement and Extension Projects. This suggests that the problems in these two projects may embrace elements attributable to the target country itself, rather than to a given sector. Nonetheless, no particular problem was raised about the Project on Electric Power Technical Standard Establishment in Laos since the government has declared its commitment. Partly because of the limitation of information based on the primary evaluations, it is not necessarily appropriate to compare these projects. Still, the commitment of the government may have affected the differences in scores by sector since the government's priority is related to financial sustainability. However, it must be noted that the priorities of a government are not always consistent with development needs.

3) Differences in Perspectives between External Evaluations and Internal Evaluations

Project evaluation results based on the information given in reports were compared in order to examine differences between external evaluations and internal evaluations, as in the quality of primary evaluation. Figure 4-13 shows the mean scores given by external evaluators and internal evaluations by evaluation criterion. Table 4-10 shows the result of the statistical analysis of the differences between external and internal evaluators, using the same method as that used for the quality of evaluations.

As these Figure and Table show, internal evaluators generally tend to be stricter in evaluating projects based on reports. The difference in the mean scores given by the two groups of evaluators is particularly large for "relevance" and "effectiveness," where internal evaluators gave some harsh scores. However, the scores for "efficiency" and "impact" do not exhibit much difference between those given by the internal evaluators and those given by the external evaluators. This corresponds to observances made in the secondary evaluation on the quality of the primary evaluations: (1) external evaluators gave lower ratings for "impact" among the DAC Five Evaluation Criteria than did internal evaluators, and (2) there was less of a difference in the scores for "efficiency" between the two groups of evaluators compared with the rest of the evaluation criteria.

Figure 4-13 Mean Scores by Evaluation Criterion

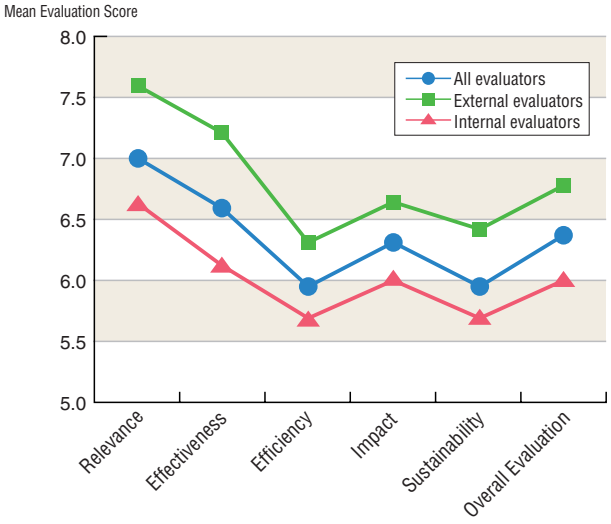


Table 4-10 Examination Results on the Differences between External and Internal Evaluations

Evaluation criteria	Differences in the population mean
Project evaluation observed from the reports (Five Evaluation Criteria)	
Relevance	○ (External)
Effectiveness	○ (External)
Efficiency	○ (External)
Impact	○ (External)
Sustainability	○ (External)
Overall evaluation	○ (External)

○ indicates a case where it has been inferred that there are differences in the mean scores of the population; whereas – indicates a case where it has been inferred that that there are no differences.
(External) indicates a question for which external evaluators are determined to have given higher mean scores.

3 Summary of Evaluation Results and Recommendations

(1) Quality of the Primary Evaluation

It is fair to conclude from the results of the secondary evaluation that JICA's terminal evaluations fulfill a certain level of quality on the whole. Year-to-year improvements in quality are also observed; the quality of "data collection," and "assessment of performance" and "analysis" are steadily on the rise. On the other hand, the quality of "recommendations/lessons learned" is relatively low and has shown little improvement. "Recommendations/lessons learned" are important for the improvement of projects through the utilization of evaluation results, and therefore more efforts are required to improve quality.

As far as evaluation methods are concerned, it is necessary to collect information not only from parties and agencies involved in projects, but also from beneficiaries while encouraging greater participation from partner countries in evaluations. It is also recommended to collect sufficient qualitative and quantitative data and to conduct more in-depth analysis by combining several analysis methods. In this connection, the quantitative aspects, such as collection of quantitative data and application of quantitative analysis methods, are weak. It is also generally observed that performance assessment in terms of achievements of overall goals is no more than a mere formality. One of the reasons for this may be the fact that the occurrence of effects of a project on the overall goal is limited at the time of the project's completion. Nevertheless, the question of whether these effects occur or not determines the value of a project as long as the project is carried out to overcome development challenges, which can be evaluated by assessing the achievement of its overall goal.

Furthermore, in order to perform appropriate evaluations, the project itself plays a significant role in providing appropriate purposes and indicators as well as in appropriately establishing a logical framework for achieving effects. It is difficult to evaluate a project if its evaluability is low with insufficient plans and data, and consequently, the quality of the evaluation tends to be low. Naturally, a project with low evaluability does not always lead to a low quality evaluation and vice versa. However, as the four terminal evaluations chosen as good evaluations by the secondary evaluation as well as the results of the secondary evaluation on these projects in question show, proper project planning, monitoring, and management are crucial for carrying out high quality evaluations. High quality evaluations are important so that readers of the reports can duly appreciate those projects that have favorable outcomes.

Among the primary evaluations subject to the secondary evaluation, some evaluations were rated high for each criterion of "evaluation framework," "data collection," "analysis/evaluation," "recommendations/lessons learned," and "reporting," or for several other criteria. They can be role models like the aforementioned four project evaluations. The belief is that sharing knowledge on how these high quality evaluations were performed and the evaluation results compiled is useful for improving the quality of evaluations. For example, not a few reports demonstrated insufficient writing skills; however, such a problem could easily be solved by referring to those reports where the presentation quality is good. In addition, many reports are compiled on the assumption that they will be used mainly as operational documents by concerned parties, and the aim of explaining the projects to outside readers has not been emphasized. Thus it seems necessary to establish a common recognition of this issue.

(2) Project Evaluation Based on the Information Given in Reports

It is fair to conclude that "relevance" of the target projects was generally high, and the results attained a certain level of outcome. While some projects achieved a high level of outcome, a few received low evaluation results; and the scores for "sustainability" vary widely from project to project. As for "sustainability," the first secondary evaluation that targeted 40 terminal evaluations in fiscal 2001 also concluded that there were substantial differences among the projects. In order to upgrade the sustainability of JICA's projects as a whole, it seems useful to first identify projects with high sustainability using the results of secondary evaluation, as well as analyze the experiences of such projects and share the results as an asset of the entire organization for the spread of good practice.

"Efficiency" was rated slightly lower than other criteria in general. The issue of timing in terms of inputting human resources and equipment was pointed out in the evaluators' comments on "efficiency." Many expressed that it was difficult to identify highly efficient projects because the economic efficiency and/or cost-effectiveness of the input are not described sufficiently. The project evaluation based on the information given in the reports of the first secondary evaluation also rated efficiency low for the same reason. Insufficient evaluation from the viewpoint of cost was indicated as an issue associated with efficiency in the primary evaluation on quality as well. Incorporating cost aspects into a project management is one of the issues to be considered in planning, implementation, and evaluation of a project.

(3) Improvement of the Secondary Evaluation Check Sheet

The Secondary Evaluation Working Group worked to recommend improvements for the secondary evaluation check sheet; specifically, recommendations on weightings of the sub-criteria for each evaluation criterion and suggestions for improvements concerning criteria.

To this end, the working group analyzed which sub-criteria were emphasized for evaluations and formulated a plan for a statistical analysis. Along with this, the sub-criteria that have such strong correlations that similar information can be obtained even when combined were statistically analyzed, and a proposal to reduce and combine these sub-criteria was drawn up*.

Using this proposal as a reference, and based on the results of the secondary evaluation, the working group further discussed the criteria and the weightings to be applied in order to improve the quality of evaluations in the future before finalizing the proposal of the improved check sheet. The improvement proposal is shown in Table 4-11. In order to compile a design for the weightings, the sub-criteria were assigned up to five points each, and the points for major criteria (the sum of all the points of the sub-criteria) were calculated. Then the weightings of sub-criteria for each major criterion were examined and the points for major criteria were allocated to each sub-criterion according to its weighting. Since the number of sub-criteria is different for each major criterion, the points given to each major criterion are not the same. As described in section 1-1 (4), Evaluation Methods, the total score, which is the sum of the points of the major criteria, does not directly indicate the quality of evaluation, and therefore, the section of the total score is not provided. Thus, the weighting of each major criterion is not established.

Finally, some points that were noticed during the process of this study are described below for the implementation of future secondary evaluations. First, secondary evaluation is limited because the evaluation is based on primary evaluation reports. Thus, it is difficult to make proper judgments unless the reports contain sufficient and relevant information. The information required for secondary evaluation is necessary for not only the secondary evaluation but also for third parties to determine whether the target primary evaluation was conducted appropriately. In this sense, the aforementioned information on the composition of the study team and evaluation methods needs to be described in the report. In addition, the criteria directly associated with the project itself, such as "evaluability," are more difficult to assess than are other criteria.

Furthermore, this year's secondary evaluation carried out several trial evaluations, including the design of the check sheet used for the actual secondary evaluation, with the participation of all the members using the same report. During the process, a common awareness of the significance in the perspectives of evaluation and focus points of sub-criteria was deepened and the diversity of evaluation results among evaluators was standardized. When a secondary evaluation is performed by different evaluators, it is important to share awareness of the evaluation perspectives among evaluators, and thus trial evaluations are useful.

Table 4-11 Improvement Proposal for the Secondary Evaluation Check Sheet

Major Criteria	Sub-criteria	Working Group's Proposal
Evaluability	Evaluability of the initially prepared PDM	4
	Evaluability of outputs, project purpose and overall goal	4
	Logic of project design	6
	Project monitoring	6
Evaluation Framework	Time frame of evaluation study	3
	Evaluation team composition- Impartiality	6
	Evaluation team composition -Specialty	
	Level of counterpart participation	6
Data Collection	Evaluation questions	4
	Data collection methods	8
	Data/information sources	
	Appropriateness of data sources	2
	Sufficiency of data/Information obtained	6
Assessment of Performance	Measurement of results	4
	Examination of project implementation process	4
	Examination of causal relationship—Logic of project design	6
	Examination of causal relationship—Before and after	6
Analysis	Objectivity of analysis	5
	Holistic analysis	6
	Analysis of promoting and impeding factors	4
Evaluation	Relevance	5
	Effectiveness	5
	Efficiency	5
	Impact	3
	Sustainability	5
	Basis of evaluation results	0
	Conclusion	7
Recommendations /Lessons Learned	Relevance of recommendations	6
	Convincing recommendations	
	Sufficiency of recommendation	4
	Usability of recommendations	5
	Relevance of lessons learned	6
	Convincing lessons learned	
	Sufficiency of lessons learned	4
	Usability of lessons learned	5
Reporting	Presentation/ legibility	7
	Clarity	
	Utilization of tables and figures	4
	Presentation of primary data	4
General Criteria for Good Evaluation	Usefulness of evaluation information	5
	Impartiality and independence	5
	Credibility	5
	Participation of the partner country	5

*See Technical Notes at the end of Chapter 2 (*7 Statistical analysis concerning proposed weightings of sub-criteria and *8 Selection of evaluation criteria).

Chapter 2 Improving JICA's Evaluations and Projects (Recommendations)

Hiromitsu Muta
Chairperson of the Secondary Evaluation Working Group
Chairperson of the Advisory Committee on Evaluation

[New Attempts in Secondary Evaluation]

The result of the second secondary evaluation has been published, following the first one in fiscal 2003. In the last fiscal year, it was revealed that each external expert had evaluation tendencies, and it was specifically indicated that it was important to compare evaluations performed by several evaluators in order to obtain bias-free evaluation results. New attempts were made this fiscal year, such as an analysis of year-to-year changes by comparing data with the last fiscal year and the participation of JICA staff in the secondary evaluation, in addition to the participation of external experts.

If the evaluation results are not reliable, the feedback effects will be limited. Even though there may be no evaluation results that convince everyone, it is still possible to build a framework to obtain an evaluation result that will convince as many people as possible. The secondary evaluation was devised to be such a framework.

Comparisons between JICA staff and external experts are not just interesting. They are also beneficial. Although the results for this fiscal year have not clarified which viewpoints differ between the two groups and why, we still need to continue these attempts. JICA is endowed with highly competent human resources. It is primarily necessary to utilize those human resources within JICA for evaluations.

The secondary evaluation activities carried out by JICA staff with external experts would enable JICA staff to understand the evaluation from a third-party perspective, and at the same time allow external evaluators to develop evaluation views without self-righteousness or bias. In evaluation, the differences between external experts and JICA staff rest merely in their respective positions and experiences, and there can be no claim of superiority on the part of either party. Mutual understanding of each other's viewpoints will nourish the evaluation capacity of individual evaluators and at the same time substantiate evaluation.

[Evaluation of Evaluators]

Identification of the evaluation tendencies of evaluators means evaluation of evaluators. In general, certain evaluation tendencies of evaluators—not limited to experts—are not always harmful in nature. It is natural that the evaluators' specialties, experiences, and sense of value influence evaluation tendencies. Nonetheless, it is still desirable to obtain the same evaluation results whenever the same evaluator performs. If the result changes every time, it is not reliable. It is not certain whether this actually happens since, in reality, one person does not evaluate the same project many times. We can measure reliability by calculating general evaluation tendencies and dividing the evaluation tendencies into true tendencies and errors by way of the method used in this study. If the errors are significant and the evaluation tendencies fluctuate widely, the evaluation results are not reliable. It would be possible to improve the qualification of evaluators through the continuation of such attempts.

[Field-oriented Approach of Evaluation]

The new ODA Charter and the new ODA Medium-term Policies emphasize a field-oriented approach. The field-oriented approach should include evaluation activities, in addition to formulation and implementation of projects. If most evaluations are lead by overseas offices in a decentralized way, a framework should be established to check and maintain quality at a certain point. The more important evaluations become, the more necessary quality assurance becomes. Using this secondary evaluation scheme, it will be possible to establish a framework to check the quality of evaluations at the headquarters. If there is any problem with the quality of evaluations, it will also be possible to identify the cause, whether it is the projects, evaluators or the implementation structures of the overseas offices.

[Emphasis on Outcomes]

Important criteria for determining the success of a project is the occurrence of outcomes, which is the evaluation viewpoint that generates large differences in judgements between external experts and JICA staff in the results of secondary evaluations. In the context of a PDM, it is the viewpoint that determines whether or not the overall goal has been achieved. Without outputs no outcomes will occur; however, even if there are outputs, outcomes do not always spontaneously occur. That is partly because it takes time for outcomes to occur, but more importantly, it is because the pathway from outputs to outcomes is relatively narrow. The number of outputs that contribute to the outcome is not usually one, and the

degree of contribution varies. If the degree of contribution is originally small, no visible changes will occur in outcomes even after time passes. Besides, the more time that passes, the stronger the influence of outputs unrelated to the project becomes. Thus it is difficult to verify whether the change in outcome is caused by the output of the project, even if there is a change. It is theoretically understood that outcomes will occur as time passes, but the chance to actually measure the contribution gets smaller as time passes.

A project generally lasts three to five years. At the time of the completion of a project, we expect an incipient outcome resulting from the activities conducted at the initial stage of the project, and it is important to identify this emergence accurately. Even if there is a long time period between an output and the emergence of a full-scale outcome, it is possible to provide a more accurate outlook of the occurrence of outcomes by citing various intermediate outcome indicators on its way. Instead of describing in the report, "Outcomes have yet to occur," one should make an effort to describe in detail what kind of outcomes are starting to take shape.

[Evaluation with Program-oriented Viewpoint]

It was considered important to carry out a project with the prospect of a causal relationship: the sequence from project activity to output to outcome to impact. And it was considered desirable to evaluate the project in that order. These ideas still hold water today, but we can have more than one pathway if the cause-effect relationship is reversed, moving backwards from what kinds of outcomes are needed to get impacts, and then, what kinds of outputs are necessary to obtain outcomes, and so on. Based on the outcome-focused trend, it is first necessary to specify what kinds of impacts and outcomes are needed. In order to significantly contribute to the occurrence of such impacts and outcomes, the formulation of a project requires a program-oriented viewpoint that allows for systematic projections of various elements. Thus, when a program is designed to achieve impacts by means of the multiple sequences of cause and effect from a program-oriented viewpoint, the same viewpoint is necessary for evaluations as well.

There are also many elements that have previously been clustered in one category of external factors in the PDM, but that should naturally be addressed as part of a project. If that is the case, it is important to carry out evaluations while giving due consideration to various elements from the program-oriented viewpoint.

[Tools for Project Improvement]

Evaluation requires money, so methods for improvement and how much improvement has been made must be considered in terms of the evaluation's expense. The primary objective of evaluation is to make improvement. This analysis showed improvements in the quality of evaluation reports, which are attributed to recent efforts, including the compilation and thorough implementation of evaluation guidelines. JICA also introduced a consistent evaluation system from the ex-ante to ex-post stages in 2001, and revised its guidelines in February 2004. It is worth noting how these efforts are reflected on the quality of evaluation.

Paying attention to the evaluation viewpoints used in this report can generate guidelines as to how a project should be formulated to obtain outcomes easily. If a project plan is based on the understanding of what is being evaluated and on attention to achievements, the quality of project formulation will be improved accordingly. The secondary evaluation sheet developed by this working group can also be used as a check list during project formulation. Since one understands the perspectives from which a project is consequently evaluated, it is easier to attain outcomes when a project is formulated with due consideration given to these perspectives. Similarly, it is better to improve the outcomes of the project itself and to verify the improvement by presenting, in an easy way, the points to be considered when implementing a project.

If a report forces you to read between the lines to find points for improvement, then the report is not practical at all. It will be increasingly important in the future to formulate guidelines and check sheets for formulation and implementation of projects based on evaluation reports, and to process and transmit information so that anyone can use evaluation results easily.

[Closing]

Evaluation is a mechanism for providing quality assurance of projects. It goes without saying that every staff member's commitment is important for improving the quality of projects, but it is also important to establish a mechanism that ensures this quality and supports their commitment. The evaluation activity presented in this report is still in the second year, and we want to see JICA exert efforts to continue ambitious evaluation activities such as those presented above.

Reflecting on the Results of Fiscal 2004 Secondary Evaluation

—Usefulness of the Secondary Evaluation Results and Application by JICA

Chairperson of the Evaluation Study Committee, JICA
Seiji Kojima, Vice-President

I would like to express my sincere gratitude to the members of the Secondary Evaluation Working Group of the Advisory Committee on Evaluation who examined better methods and implemented reliable and convincing evaluations on 58 terminal evaluations using their expertise.

JICA set up the Advisory Committee on Evaluation in fiscal 2002 and introduced the secondary evaluation performed by the committee members. The secondary evaluation is considered significant for JICA to increase the transparency and objectivity of its evaluations and to provide more proper accountability. It is also quite beneficial to hear suggestions for improvements in the quality of evaluation. Consequently, the result of the secondary evaluation conducted in the last fiscal year was compiled and published in *the Annual Evaluation Report 2003*. The JICA Guideline for Project Evaluation was then greatly revised in February 2004 utilizing valuable suggestions and advice. JICA is currently making efforts to consolidate the evaluations based on the revised guidelines and improve the quality of evaluations.

The secondary evaluation in this fiscal year has provided us with a large number of valuable suggestions concerning the quality of evaluations and JICA projects through more in-depth analysis. The evaluation result states that evaluations and projects have achieved a certain level of quality, and that the quality of evaluation in particular is rising every year. This is good news for JICA, which has been making efforts to improve evaluations and projects. However, it also indicates that various issues remain, such as the low quality of recommendations/lessons for feedback into new projects, the weak perspective of efficiency and impact, the participation of developing countries in evaluation, and insufficient information collection from beneficiaries. Based on these suggestions, JICA will continue in its efforts to improve evaluations and projects.

As part of these efforts, as indicated in the recommendations, JICA will continue and enhance secondary evaluation in the future, and will make efforts to improve the quality of primary evaluations using the secondary evaluation check sheet presented in this report for better management of JICA's evaluations.

JICA is currently promoting a field-oriented approach toward the implementation of more effective projects, and overseas offices are increasingly playing a more central role in the identification, formulation, planning, implementation, monitoring, and evaluation of projects. Under such circumstances, enhancement of the evaluation system and improvement of the quality of evaluations of overseas offices are important issues for promoting the field-oriented approach. Since secondary evaluation helps overseas offices control the quality of evaluations and improve projects through the evaluations, we will examine the systematic use of secondary evaluation.

In addition, awareness of the evaluation viewpoints during project formulation is essential for the outcome-oriented management of projects and, accordingly, JICA will consider the use of the secondary evaluation check sheet from that point of view.

In the meantime, in light of the importance of a program-oriented viewpoint in the formulation and evaluation of projects toward the solid achievement of results and impacts, JICA is currently committed to enhancing its program approach. To be more specific, in fiscal 2004, JICA introduced a system to manage and operate the planning and implementation of projects within the program unit, which had previously never been carried out beyond mere conceptualization, using budget allocations at the program level. As for evaluations, based on experiences in country-program evaluations and program-level evaluations, JICA is currently making efforts to improve evaluation methods and programs using evaluations in order to respond to the formation of programs and improve the quality of evaluations of programs. JICA, with advice from the Advisory Committee on Evaluation, will continue our efforts in these tasks.

Finally, in response to advice stated in the secondary evaluation of last fiscal year that it is useful to present specific examples of good evaluation to improve the quality of evaluations, JICA launched the JICA Good Practice Evaluation Award in fiscal 2004. By selecting evaluations demonstrating good feedback of results and quality evaluations as good practice and widely sharing them as role models, JICA intends to learn at the organizational level. Four terminal evaluation reports that were regarded as good evaluations in the secondary evaluation were selected and awarded Outstanding Evaluation Awards as part of the first JICA Good Practice Evaluation Award. JICA would like to accumulate positive evaluation models that will lead to improvements in the quality of evaluations by selecting high quality evaluations objectively using the results of secondary evaluation and sharing the characteristics of good evaluations through JICA Good Practice Evaluation Award.