Part 3 External Evaluation

–Secondary Evaluation by the Advisory Committee on Evaluation

Part 3 External Evaluation
–Secondary Evaluation by the Advisory Committee on Evaluation

In FY 2002, JICA established an Advisory Committee on Evaluation in an effort to improve its evaluation system and methods with the help of external experts and improve the objectivity of evaluations by having them examined by these experts. The committee members, which include academics, NGO members, and journalists, are independent of JICA and experts on development aid and evaluation. (See Part 1, Chapter 2, 2-5 "Promoting Evaluation by Third Party")

Since their first meetings in June 2002, the committee members have met roughly every two months to assess JICA's evaluations (a process called secondary evaluation) and discuss possible ways to improve evaluation methods as well as how to best use evaluation results to improve project planning and implementation. The Committee also provides JICA with advice and suggestions concerning a wide range of issues, including problems related to enhancing JICA's evaluation system, evaluation approaches for newly targeted cooperation schemes, and how to improve disclosure of evaluation findings.

As part of JICA's effort to ensure objectivity and transparency in evaluation and improve the overall quality of evaluation through third party examination, JICA asked the Committee to conduct secondary evaluation to assess the terminal evaluations of, and point out issues for discussion for, the 40 Project-type Technical Cooperation projects carried out in FY 2001. See Chapter 1 for the full text of the secondary evaluations.

The Committee's secondary evaluation revealed a whole range of problems with JICA's evaluation, including those JICA has already recognized and those discovered for the first time owing to the external experts' independent input. JICA plans to initiate concrete efforts to improve the quality of its evaluations based on the secondary evaluation results and the suggestions made by the committee members (This is discussed in Chapter 2, "JICA's Response to the Secondary Evaluation Results by the Advisory Commitee on Evaluation"). JICA also intends to continue developing the use of secondary evaluation.

Study Team for the Secondary Evaluation:

Chairperson of the Advisory Committee on Evaluation:

Hiromitsu MUTA	PhD. Professor, Director of the Center for Research and Development of Educational Technology, Tokyo Institute of Technology

Committee Members:

Atsuko AOYAMA	M.D., PhD. Professor, Department of International Health, School of Medicine, Nagoya University
Kiyoko IKEGAMI	Director, UNFPA Tokyo Office
Michiya KUMAOKA	President, Japan International Volunteer Center
Tsuneo SUGISHITA	Professor, Faculty of Humanities, Ibaraki University
Masafumi NAGAO	Professor, Center for the Study of International Cooperation in Education, Hiroshima University
Shunichi FURUKAWA	PhD. Professor, Institute of Policy and Planning Sciences, University of Tsukuba
Atsushi YAMAKOSHI	Manager, Trade and Investment Policy Group, International Economic Affairs Bureau, Japan Business Federation

Chapter 1 ● Results of the Secondary Evaluation by the Advisory Committee on Evaluation

Chairperson, Hiromitsu MUTA

1-1 Objectives of the Secondary Evaluation

The secondary evaluation by external experts of JICA's terminal evaluations should serve the following two objectives.

(1) Securing Transparency in Evaluation Procedures

JICA performs a terminal evaluation of all its projects. In terms of transparency of these evaluations, however, are more or less internal. Even though they include experts and consultants as evaluation team members, most have been either a member of a supporting committee in Japan or involved in the project in some other way, such as providing advice or services. While people involved in a project can produce more informative evaluations than outsiders because of their better knowledge about the project and its circumstances, their evaluations might lack objectivity.

Even external evaluations by a third party, however, are not necessarily more objective because a third party might have an interest in JICA or in the project they are evaluating. More importantly, opinions are inevitably influenced by personal views.

These problems can be solved by having members of the Advisory Committee on Evaluations consist of external experts who evaluate the results of the internal primary evaluations.

(2) Securing Credibility of Evaluation Results

A secondary evaluation can ensure that evaluation results are reliable and unbiased. The terminal evaluation of a project is performed only once by a evaluation team dispatched to the project site. But sending another evaluation team is not likely to produce identical findings. It is, of course, virtually impossible to send several different evaluation teams to a project site for one project and then compare their findings. One possible means of deriving reliable conclusions from the results of a single evaluation made by a specific group of people is to have the findings examined by several other evaluator in a secondary evaluation. While secondary evaluation inevitably has limitations because they are based on a primary evaluation, they nonetheless can be expected to produce less biased and more reliable conclusions by the involvement of several evaluators. This is as true for external evaluation as well as for internal ones.

Also important is the fact that a secondary evaluation is usually performed about a year after a primary evaluation. This increases the chance that a second evaluation will lead to new findings, such as new evidence showing the effect of the project.

1-2 Secondary Evaluation Procedure

(1) Focus of the Secondary Evaluation

The secondary evaluation focused on two things. First, it looked at the evaluation method and the quality of the evaluation reports, whether the evaluation report contained enough information and whether the evaluation method was appropriate. Second, the secondary evaluation focused on the performance of the project as described in the evaluation reports. These two foci are closely linked to one another. This study examines the 40 terminal evaluations conducted in FY 2001. This particular secondary evaluation applied a meta-evaluation method.

(2) Ratings by Evaluation Categories

The Advisory Committee on Evaluation composed of external experts first read the terminal evaluation reports and then conducted a secondary evaluation. First, evaluation reports were rated within 27 categories (8 on the evaluation framework, 6 on how the evaluation study was performed, 9 on information analysis and evaluation, and 4 on the quality of lessons and recommendations). Then, a secondary evaluation on the performance of the projects themselves was conducted in 6 categories and on a 1-to-5

scale. These categories and the standards for the secondary evaluation described in the evaluation sheet are based on "The Standards for Good Evaluations" in the JICA Evaluation Guidelines.

If each committee member had read all the evaluation reports and made a secondary evaluation individually, the scores obtained would reflect the opinions of all the members without the bias of any particular member simply by calculating the average of their ratings for each category. As each committee member has a different position and opinion concerning specific issues, to secure impartiality in a secondary evaluation it is necessary to obtain the ratings made by several members and average the scores. This was not possible, however, because time constraints made it impractical to ask every member to evaluate all 40 reports. So the Committee chairperson evaluated all the reports while the other members dealt with 6 to 9 reports each to ensure that each report was evaluated by at least two members. Since only two or three committee members evaluated each report, simply averaging out their ratings cannot reduce individual bias sufficiently to ensure the credibility and impartiality of the secondary evaluation.

In theory, the ratings of Committee members can be broken down into real ratings (ratings not affected by personal bias) and the coefficients of each member's evaluation tendencies. Therefore, statistical analysis is applied to separate these two factors and adjust for personal tendencies, such as differences in generosity when rating a certain category. This way, the personal biases that distort results are reduced to a satisfactory extent to produce impartial ratings.

(3) Estimated Value of Ratings by Category

For all 40 reports, we computed for each category the estimated value of rated scores (real ratings) and each evaluator's tendency coefficient. By adjusting the tendency coefficients for each category so that their total sum is zero, we ensured that each estimated value of rated scores equals the average of the rated scores that would have been produced had every committee member evaluated all the reports.

Table 3-1 shows the averages and standard deviations of the estimated value of rated scores by category for all 40 reports using the calculation method described above. Both the figures for the 33 small categories and values for larger category groups are given.

(4) Individual Member Evaluation Tendencies

The coefficients for individual member evaluation tendencies showed various patterns. Some tended to give high ratings in all categories, while others showed a general tendency to give low ratings. Still others tended to give high ratings or low ratings only in certain categories. The findings demonstrated that evaluation methods like a five-scale rating system is not free of personal bias, underscoring the importance of obtaining averages from as many evaluators as possible in order to minimize bias. No significant correlation was found between the tendency coefficients of the evaluators, meaning that their evaluation tendencies were independent of each other.

In addition, the standard deviations of individual evaluator tendency coefficients were calculated for each category and shown in Table 3-1. For evaluation of evaluation report quality, the standard deviations ranged from 0.13 ("Credibility" of the "Study process") to 0.67 (the "Usefulness" of the "Lessons"). The values for some categories are greater than the standard deviation of the scores for all 40 categories. A larger standard deviation value indicates wider differences among evaluator opinions. In other words, one individual's evaluations are intrinsically unreliable and inconsistent. This is deduced from the variance of rated scores among different evaluators for some evaluation categories being greater than the variance of scores among different projects.

Conversely, scores rated for project performance based on the information on their reports of the primary evaluations showed relatively smaller differences in opinion among the evaluators. The standard deviations of the rated scores for project performance based on the evaluation reports ranged from 0.22 (Relevance) to 0.50 (Effectiveness). Their ratings diverge considerably for such categories as "Impact" and "Overall score". This is partly because there is not yet a widely accepted method for measuring and evaluating the social impact of a project.

While differences in opinion are inevitable due to divergence in values, it is likely that vague definitions of words and unclear judgment criteria are also partly responsible. Previous studies have shown, however, that differences due to such factors can be reduced to some extent by making definitions and judgment criteria consistent. For example, thorough discussion among all the evaluators over a small number of actual evaluations establishes fully consistent judgment criteria before carrying out a large number of evaluations.

Table 3-1 Average Scores and Standard Deviations for Individual Evaluation Categories

( ) Standard Deviations for Reports
(Italicized Figures) Standard Deviations for Evaluator Evaluation Tendencies

I. Evaluation Research / Evaluation of Reports

Categories	Usefulness	Fairness / Neutrality	Credibility	Participation of Recipient Country Side	Overall Score
1: Evaluation Framework					
●Timing, Duration of Evaluation Study	3.24 (0.40) *(0.57)*				
●Evaluators / Member Composition		2.69 (0.27) *(0.55)*	3.12 (0.30) *(0.58)*	3.18 (0.57) *(0.47)*	
●Study Process	3.16 (0.51) *(0.26)*	3.07 (0.39) *(0.37)*	3.12 (0.53) *(0.13)*		
●Study Cost (Appropriateness of Scale of the study)	2.96 (0.40) *(0.24)*				
Total	3.12 (0.45) *(0.40)*	2.88 (0.38) *(0.46)*	3.12 (0.43) *(0.42)*	3.18 (0.57) *(0.47)*	3.07 (0.46) *(0.46)*
2: Implementation of Study					
●Items Studied (Terms of References)	3.25 (0.49) *(0.39)*				
●Information Collection Methods	3.16 (0.55) *(0.43)*	2.86 (0.47) *(0.42)*	3.10 (0.51) *(0.14)*		
●Visited Sites / Interviewee Composition		2.80 (0.38) *(0.48)*		3.28 (0.47) *(0.29)*	
Total	3.20 (0.52) *(0.41)*	2.83 (0.43) *(0.45)*	3.10 (0.51) *(0.14)*	3.28 (0.47) *(0.29)*	3.08 (0.51) *(0.40)*
3: Information Analysis / Evaluation					
●Information Processing / Analysis	3.13 (0.75) *(0.39)*	3.00 (0.44) *(0.54)*	3.23 (0.65) *(0.25)*	3.23 (0.50) *(0.32)*	3.15 (0.60) *(0.40)*
●Results of Evaluation on Five Evaluation Criteria					
1) Relevance	3.45 (0.60) *(0.36)*				
2) Effectiveness	3.25 (0.56) *(0.41)*				
3) Efficiency	2.76 (0.63) *(0.56)*				
4) Impact	3.33 (0.64) *(0.40)*				
5) Sustainability	3.45 (0.56) *(0.46)*				
Total	3.25 (0.65) *(0.44)*				3.25 (0.65) *(0.44)*
●Quality of Lessons / Recommendations					
1) Lessons	3.14 (0.61) *(0.67)*				
2) Recommendations	3.21 (0.61) *(0.44)*	3.01 (0.38) *(0.46)*	3.12 (0.46) *(0.50)*		
Total	3.18 (0.61) *(0.57)*	3.01 (0.38) *(0.46)*	3.12 (0.46) *(0.50)*		3.12 (0.53) *(0.56)*

II. Evaluation on Projects' Performance based on the Information given in Reports

Criteria	Evaluation on 1-5 Scale				Overall Score
● Relevance	3.49 (0.67) *(0.22)*				
● Effectiveness	3.27 (0.66) *(0.37)*				
● Efficiency	2.90 (0.65) *(0.39)*				
● Impact	3.23 (0.68) *(0.50)*				
● Sustainability	3.12 (0.87) *(0.25)*				
Total	3.20 (0.74) *(0.36)*				3.20 (0.74) *(0.36)*
● Overall Rating	3.21 (0.71) *(0.47)*				3.21 (0.71) *(0.47)*

Note: Rated on a 1 to 5 scale (1: low, 2: rather low, 3: average, 4: fairly high, 5: high)

The ratings for the "Usefulness" of the "Lessons" for the categories for evaluation on report quality showed especially large dispersion by evaluators. Differences in opinion among the evaluators were also great for the "Credibility" of "Recommendations". Since lessons should produce recommendations and improvements for similar projects in the future, wide differences in the rated scores for the "Usefulness" of "Lessons" by evaluators is extremely important issue. In their remarks, some expressed skepticism about the grounds for the lessons and recommedations. The evaluators' opinions varied over the most important part, the utilization of evaluation results, suggesting that analysis in the primary evaluation lacks objectivity. This underscores the need to establish consistent rules and guidelines for presenting lessons and recommendations in evaluation reports, including requirements for clarifying their supporting evidence.

(5) Rethinking the Categories for Secondary Evaluation

Among the 33 evaluation categories, many pairs show a strong correlation. This means that similar information is available from a smaller number of categories. Hence, to raise the efficiency of secondary evaluation, reducing the number of categories is an option. To ensure objectivity, categories were reduced in number by using factor analysis. Note that six evaluation categories were excluded from the factor analysis, including the "Usefulness" of "Five Evaluation Criteria", and "Lessons" and "Recommendations".

The analysis revealed that four categories – "Study Process", "Study Cost", "Items Studied", "Visited Sites / Interviewee Composition" – could be substituted with other variables. It also indicated that two evaluation criteria – "Fairness / Neutrality" and "Credibility" – could be seen as effectively the same; the degree of "Credibility" is proportional to the degree of "Fairness/Neutrality". The implication of this is that more or less similar results could be attained even if these two criteria are not evaluated separately.

As shown in Table3-2, a new set of evaluation categories is formulated by altering some of the categories, interchanging some categories with criteria, and replacing the word "Usefulness" with "Appropriateness" for some categories. As in the evaluation sheet we used, ratings based on the new secondary evaluation sheet are on a 1-to-5 scale. In addition to reducing some categories, two new categories are added that were mentioned by evaluators in their remarks about the reports – "Understandability of Evaluation Reports" and "Overall Rating of Reports".

As shown in Table 3-2, the number of categories has been reduced by a third, from the original 33 to 22. The amount of information available, however, from the evaluations will barely change. Since it is clear that this streamlined evaluation sheet using fewer categories is useful for secondary evaluations, using it for future secondary evaluations will be more efficient.

1-3 Results of the Secondary Evaluation

1) Secondary Evaluation on Evaluation Methods and the Quality of Reports

Figure 3-1(P109-111) shows the distributions of scores for the quality of the 40 evaluation reports and the secondary evaluation of the performance of the 40 projects based on the information given in their reports. As both Table 3-1 and Figure 3-1 show, the scores for the quality of evaluation reports were fairly good, with scores mostly higher than "average (3.00)" for all categories. Nonetheless, there are problems that still need to be addressed. Below, this report discusses general problems about the process of producing evaluation reports and examines by category the results of quantitative analysis in Table 3-1 and Figure 3-1, as well as committee member qualitative comments.

1) Project Design Matrix (PDM)

In many cases, the PDM was revised for terminal evaluation. In principle, terminal evaluation should be based on the most recent PDM. If the most recent PDM is unclear, it may justify a review of the PDM to clarify indicators at the time of terminal evaluation. But in a considerable number of evaluation reports, the overall goals specified in the PDM were lowered because they are too difficult to attain practically. A PDM with lowered overall goals was renamed as PDMe. This mistakenly reverses priorities. It is better to first establish overall goals and then design a project to achieve them. Typically, however, projects tended to be drawn up first, and then the overall goals were considered only on paper. In contrast, a results-based approach requires redesigning a project if that is not what is necessary to achieve the overall goals.

A project's value rests on how well it acts as a means for achieving social needs (overall goals), and not on simply executing the project which may or may not satisfy these needs.

Table 3-2 New Evaluation Sheet for Secondary Evaluation by External Experts

I. Evaluation of Evaluation Reports

Criteria / Categories	Usefulness	Fairness / Neutrality / Credibility	Appropriateness	Remarks
1. Evaluation Framework				
● Timing of Evaluation			( )	
● Evaluators / Member Composition		( )		
● Cooperation from Recipient Country			( )	
2. Information Collection / Analysis				
● Information Collection	( )	( )		
● Information Analysis			( )	
3. Analysis / Evaluation				
● Five Evaluation Criteria				
1) Relevance	( )			
2) Effectiveness	( )			
3) Efficiency	( )			
4) Impact	( )			
5) Sustainability	( )			
● Quality of Lessons / Recommendations				
1) Lessons	( )			
2) Recommendations	( )	( )		
4. Understandability of Evaluation Report		( )		
5. Overall Rating of Evaluation Report		( )		

II. Evaluation of Project Performance Based on the Information Given in Reports

Criteria / Categories	Evaluation on 1-5 Scale	Remarks
● Relevance	( )	
● Effectiveness	( )	
● Efficiency	( )	
● Impact	( )	
● Sustainability	( )	
● Overall Rating of Project	( )	

Note: Rated on a 1 to 5 scale (1: low, 2: rather low, 3: average, 4: fairly high, 5: high)

2) Reporting

While some reports are highly rated for appropriate evaluations and easy-to-understand descriptions, others are written in an unsatisfactory manner, with no data sources indicated, no attached list of interviewees, as well as information of unclear reliability. Some reports are not written logically and are therefore difficult to understand. Others provide no numerical data to support their arguments. Some reports do not contain the questions asked in interviews and questionnaires or the answers given by the interviewees and respondents. Still others lack consistent terminology, using various words for the same concepts. Obviously, more effort is needed to improve how reports are written and to train staff how to write them. Serious consideration should be given to producing a report writing manual and providing well-written reports as models.

A more fundamental problem is that many reports are written as a mere formality. These reports contain few concrete and significant comments about specific issues and problems. Reports need to delve deeply into important issues and problems. They need to give clear answers to likely questions even when read by an outsider. Reasons for poor report quality include a weak evaluation culture in Japan, lack of consideration to the parties that receive the feedback, and insufficient analysis.

3) Evaluation Framework

Scores for the "Evaluation Framework" are in large part slightly above "average (3.00)," while the score for "Fairness/Neutrality" of "Evaluators/Member Composition" is lower than for other criteria. This is because evaluation teams are mostly made up of individuals involved in the project. The scores for "Usefulness" in the "Study Cost" (appropriateness of scale) are also a little lower. Note that the number of team members and their specialty determine ratings rather than overall expense in this Study. For instance, the evaluation asks whether the evaluation team is too large for the task or whether it contains members whose presence is unnecessary given the overall composition.

The evaluation framework is solid in some evaluations, but weak in many others. The study period is often inappropriate – in some cases too long for the workload involved while in others too short. Generally speaking, the number of team members is larger than necessary. In some cases, it is quite possible for a member to carry out his own responsibility along with responsibilities assigned to another member. The study period and the number of members are the factors that directly affect study costs. The cost efficiency of evaluations needs more attention.

Since the number of evaluators also involved in the project implementation is often very high, more people not involved in the project need to participate in order to inject transparency into evaluations. In some cases, people without expertise in the area in question are involved in evaluation. In other cases, people doing evaluations have enough expertise in that specific area but lack knowledge in development assistance in general or in the country or sector. In these cases, there is no basis for comparing the project to other projects or countries. In still other cases, insufficient evaluation knowledge leads to poor or inappropriate analysis.

The timing of evaluation study needs to be more flexible. Terminal evaluation is conducted according to a predetermined timeframe, such as six months before the planned completion date. It is important to be able to respond flexibly to, for example, a delay in the delivery of certain equipment or a significant political change in the country.

4) Implementation of Evaluation Study

The scores for "Implementation of Evaluation Study" are mostly around "average (3.00)." But the scores for "Information Collection Method" and "Fairness/ Neutrality" of "Visited sites/Interviewee composition" categories are slightly lower than for other categories. It is better to use multiple methods for triangulation and select a broader range of sites for collecting information. In particular, surveys on the final beneficiaries should be enhanced.

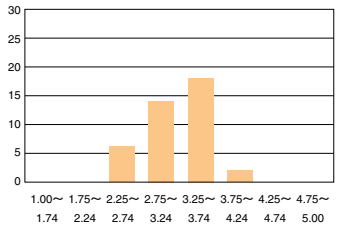
"Items Studied" is largely appropriate, and the information-gathering methods are mostly useful. Some evaluations collect information from a wide range of sources. In other reports, however, the number of interviewees was insufficient, and sources of information unclear because some reports lack an interviewee list. There are also cases where the number of direct beneficiaries covered, such as trainees and farmers, is insufficient, and cases where important sources of information, such as dispatched experts are not interviewed. It is also necessary to hear opinions from people not involved in the projects to secure objectivity.

Since partner countries are project beneficiaries, it is desirable to conduct joint evaluations with them. In fact, many evaluations are correctly carried out in this way.

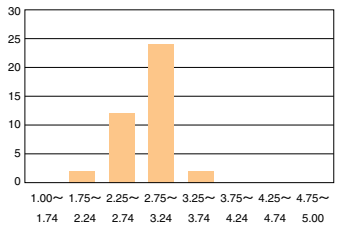
Figure 3-1 Distribution of Scores for Evaluation Report Quality and Secondary Evaluation on Projects Based on Information Given in Reports

1. Evaluation Framework

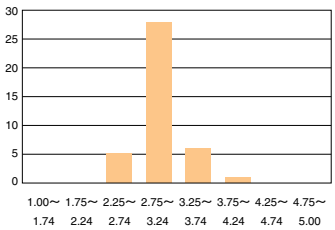
Timing of Evaluation (Usefulness)



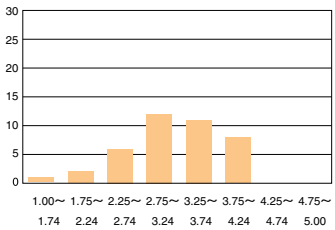
Evaluators/Member Composition (Fairness, Neutrality)



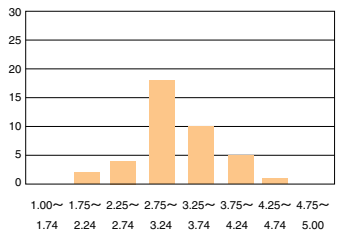
Evaluators/Member Composition (Credibility)



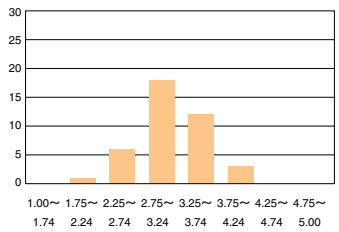
Evaluators/Member Composition (Recipient Country Side)



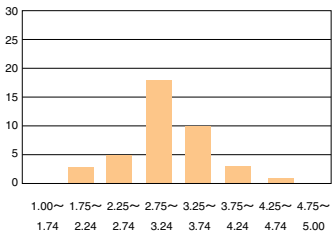
Study Process (Usefulness)



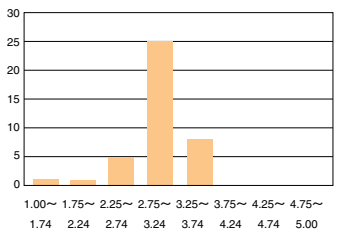
Study Process (Fairness, Neutrality)



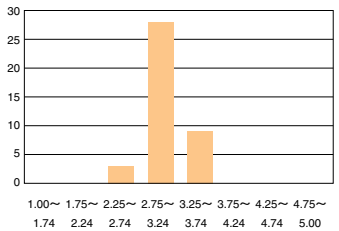
Study Process (Credibility)



Study Cost (Usefulness)

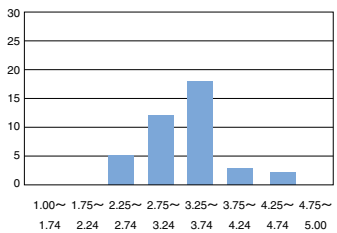


Evaluation Framework (Total)



2. Implementation of Study

Items Studied (Terms of Reference) (Usefulness)

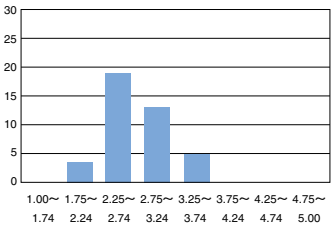


Evaluators/Member Composition (Fairness, Ne

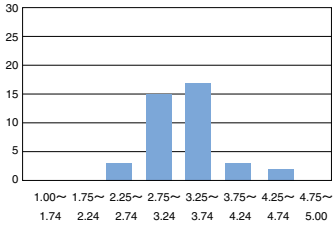
Evaluators/Member Composition (Fairness, Ne

Evaluators/Member Composition (Fairness, Ne

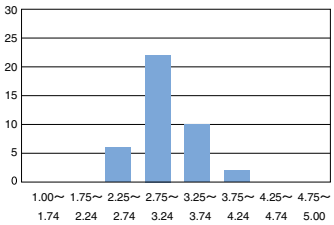
Visited Sites / Interviewee Composition (Fairness, Neutrality)



Visited Sites / Interviewee Composition (Recipient Country Side)

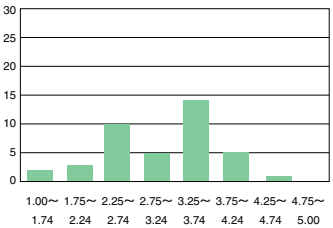


Implementation of Study (Total)

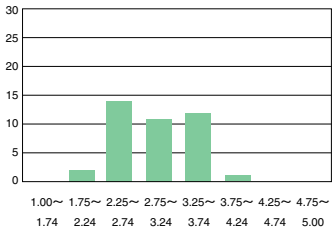


3. Information Analysis / Evaluation

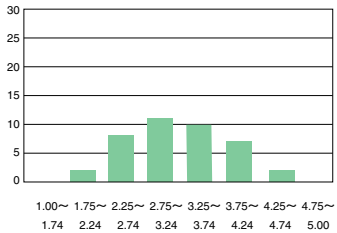
Information Processing / Analysis (Usefulness)



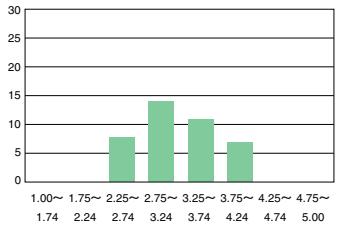
Information Processing / Analysis (Fairness, Neutrality)



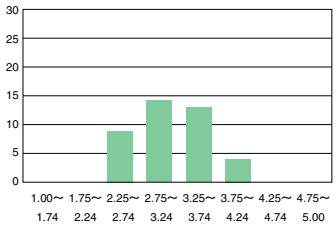
Information Processing / Analysis (Credibility)



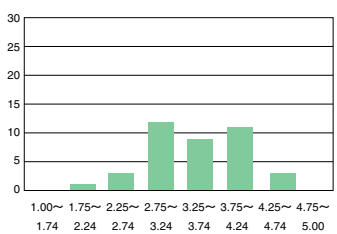
Information Processing / Analysis (Recipient Country Side)



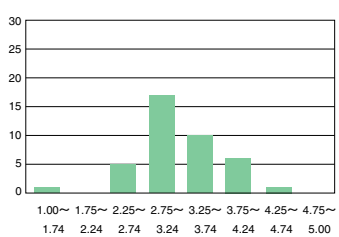
Information Processing / Analysis (Total)



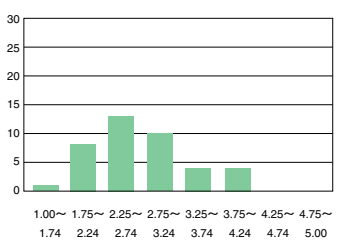
5 Evaluation Criteria: Relevance (Usefulness)



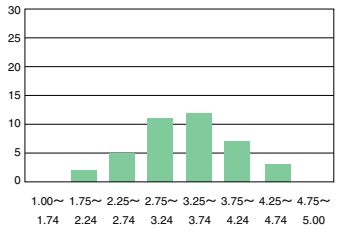
5 Evaluation Criteria: Effectiveness (Usefulness)



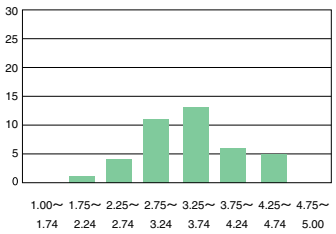
5 Evaluation Criteria: Efficiency (Usefulness)



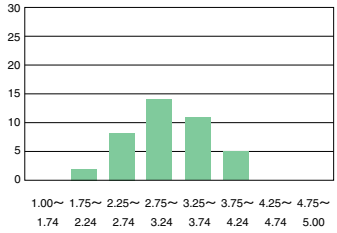
5 Evaluation Criteria: Impact (Usefulness)



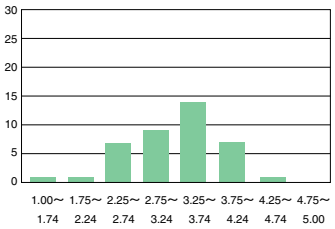
5 Evaluation Criteria: Sustainability (Usefulness)



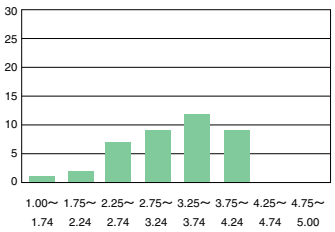
5 Evaluation Criteria: Results (Total)



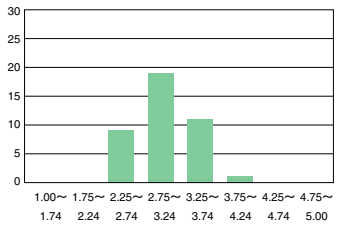
Lessons (Usefulness)



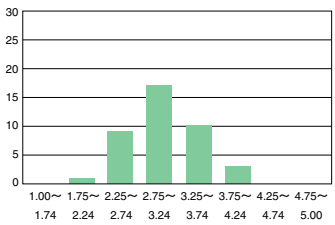
Recommendations (Usefulness)



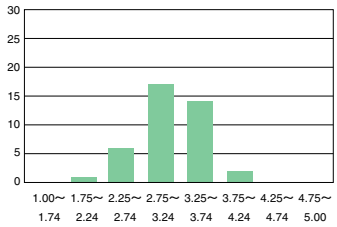
Recommendations (Fairness, Neutrality)



Recommendations (Credibility)

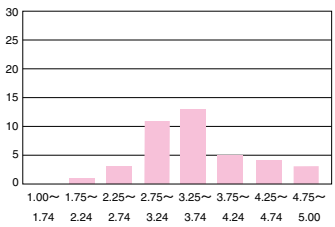


Quality of Lessons and Recommendations (Total)

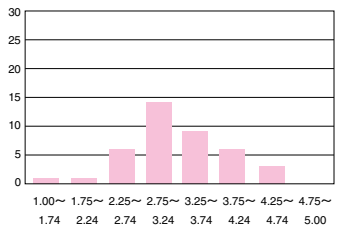


4. Evaluation of Project Performance based on the information given in Reports

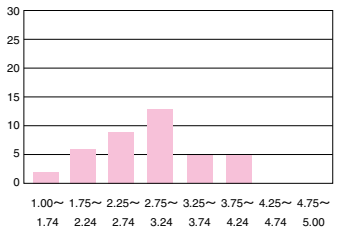
Relevance of Project



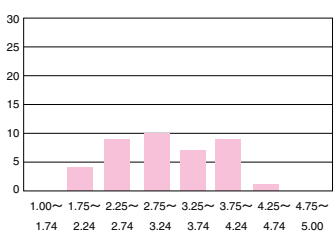
Effectiveness of Project



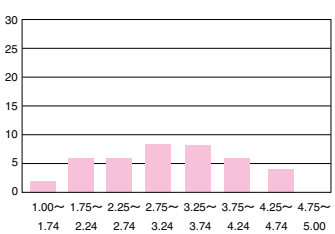
Efficiency of Project



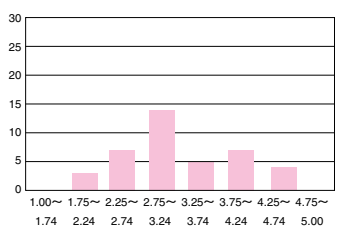
Impact of Project



Sustainability of Project



Overall Rating of Project



5) Information Analysis / Evaluation

Many reports properly analyze and evaluate information. The scores for "Information Analysis / Evaluation" are above "average (3.00)", except for "Usefulness" of the "Efficiency" category. This can be attributed to confusion about the "Efficiency" criterion. One report, for instance, used the word "efficient" to describe use of the budget as planned. Efficiency, however, means making the most of available resources including funds. Procuring necessary equipment locally at a low cost, for instance, contributes to efficiency. If equipment is purchased at a price higher than locally procured, it is impossible to judge efficiency unless the reason for the purchase (such as different spec, functions, or performance) is clear. It is necessary to establish clear criteria for determining efficiency and make these criteria known to all the evaluators.

As Figure3-1 shows, there are no large discrepancies among projects in scores for "Evaluation Framework" and "Implementation of Evaluation Study" categories. On the other hand, rather large discrepancies exist in scores concerning "Information Processing/Analysis" and "Five Evaluation Criteria." This indicates a wide difference in quality among evaluation results.

To ensure objectivity, every effort is made to use quantitative analysis, such as rating achievement on a scale. But some reports give few convincing grounds or no grounds at all, for their evaluation. Several reports have problems with data quality, using no statistical data or only very old data.

Training programs and technical transfer should be measured not only by the number of people who received or otherwise benefit from the training but also by the quality of knowledge and techniques transferred, but few evaluations deal with this issue. In many cases, insufficient attention is given to the question of how best to effectively evaluate training programs and technical transfer projects. In other cases, information is not analyzed carefully enough. In analyzing the ripple effects through society, it is important to carefully weigh economic, political, and social factors.

Several reports seem to be too generous and too favorable to the projects they evaluate.

6) Lessons and Recommendations

Some reports contain very concrete and useful lessons and recommendations, while others fail to identify good lessons or make good recommendations about the project. Some recommendations are not closely related to evaluation results, while others are too positive, or subjective and emotional. In some cases, recommendations are not concrete enough. Overly generalized lessons and recommendations are simply banal and useless. On the other hand, lessons and recommendations reflecting too many factors unique to an individual project cannot be applied to other projects.

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

Secondary evaluation results on projects based on reports description scored above average (3.00) in Five Evaluation Criteria, except in "Efficiency", and especially high in "Relevance". The reasons for low efficiency are described in the above section. Compared with the result of evaluation on report quality shown in (1), standard deviations are larger. Also, the scores greatly differ among projects. This is especially true for sustainability.

For conclusion of evaluation, most projects scored ordinary (3.00) or more, including seven projects that scored very high (4.00). Two projects scored relatively low (2.00) or lower. As described above, some projects were extremely successful projects, while a few had low evaluations.

The reasons given for unsuccessful projects were a sudden change in external conditions and improper project management. Some evaluations observed projects with low expectations due to insufficient planning or a lack of understanding as to what was intended. For these projects, it is necessary to fully verify what the original plan was.

Projects that developed problems often did not clarify their objectives in the planning stage or were very poorly designed. These seem to already have problems early in the project planning process, such as counterpart selection, appropriateness of a target for a project, inconsistency between title and project content, weak ownership of partner country, and so on.

For relevance, many evaluations superficially evaluated relevance simply from whether the project is in line with the development plan of the partner country or the Country Assistance Program of Japan. It should also be questioned if the standard of relevance is judged by whether the project is appropriate to achieving its goals. A project without a logical measure should be evaluated low in relevance. It is also important that not only Japan but also the recipient country clarify how to advance the development program for the whole recipient country and target sector, and to

judge whether Japanese assistance is appropriate to the country or sector.

1-4 Summary and Recommendations

Based on a comprehensive judgment of the above results on quantitative analysis and qualitative analysis of the remarks and comments of evaluators about the quality of evaluation reports and project performance, the following are recommendations on future terminal and secondary evaluations. The rest of this section also recommends improvements on project planning and management.

(1) Evaluability in Project Planning and Management

1) Clarification of Purposes and Goals

Some reports lack the necessary data to analyze information in the terminal evaluation. This was largely because of insufficient consideration in developing a project plan and PDM when a project is launched.

It is important to keep in mind that when formulating a project plan and a PDM to document it that these indicators are indispensable for the terminal evaluation to measure its performance. Therefore, during the planning stage of a project, it is crucial to fully consider the project purpose as well as methods for obtaining data that will be used to measure performance. When developing a PDM, participants in the recipient country need to sufficiently understand the structure and meaning of a PDM, and it is also important to carefully confirm the statistics that will be used as indicators.

It is important to set indicators numerically. If possible, set concrete numerical indicators in advance, such as the number of training, participants, development numbers, increasing rates of harvest or income, and so on, to make judging what is accomplished easier.

To set the numerical indicators for a project is, of course, important. If a project aims at a technical improvement, the project often measures what the participants accomplished with quantitative data only, for instance, number of participants. In evaluating a technical improvement, however, it is also necessary to judge qualitatively the skills that the participants acquired and whether they are useful in their daily work. It may be necessary to both compare the technical level of participants before and after training and to utilize objective quantitative indicators. When using a questionnaire survey, it is best to use a five level rating system rather than yes/no questions.

2) Improvement of Project Management by Mid-term Evaluation and Monitoring

In practice, detailed information is only available after a project has started. Even though a PDM is fully formulated before a project, it is only natural to revise it based on information obtained after a project starts. But since a PDM guides the implementation of a project, it should be revised during the formal steps of a project, such as the mid-term evaluation. In principle, terminal evaluation is done based on the most recent PDM, so in most cases it should use the PDM which revised at the mid-term evaluation. In some cases, however, the PDM was also revised for the terminal evaluation among projects examined for this secondary evaluation. This indicated that the PDM was developed as a mere formality and not used in the project.

A project should make the most of a PDM. For example, a participant, a dispatched expert, or a counterpart can use a PDM for a self-evaluation if they fully understand the PDM at the start of a project and if their own objectives are reflected in the PDM. It also enhances project efficiency. Ex-post evaluation becomes easier by implementing daily monitoring and recording important information at that time based on the PDM.

(2) Quality Improvement of Evaluation Reports

1) Member and Composition of Evaluation Teams

When organizing an evaluation team, it is necessary to select persons with high expertise and not to have a disproportionate number of specialists in a particular area. It is also necessary to consider lowering study cost by decreasing the number of members by appointing members concurrently.

High expertise here means specialized knowledge of the sector concerned, of international development issues, and of the evaluation. It is difficult to find one person with all this expertise, so what is important is to balance the evaluation team as a whole.

It is a matter of course that the recipient country participates in evaluations. Moreover, by having the evaluators from the recipient country learn PDM and evaluation methods in advance, joint evaluation is more efficient.

2) Coverage and Method on Information Gathering

In order to measure to what extent the purpose and goal is achieved, it is necessary to widely gather not only published statistics but also other primary data. For measuring impact, it is necessary to gather information from direct beneficiaries (such as farmers, participants, and patients). It is also necessary to contrive ways to evaluate understanding and thinking besides open-ended questions in the interview or questionnaire survey. It is more efficient to do a survey before the evaluation team arrives and then conduct a supplemental survey as a field study. It is also necessary to improve objectivity by increasing the number of target persons for study and distributing them widely among dispatched experts and counterparts to beneficiaries. If the primary evaluation does not obtain enough information, sufficient results from the secondary evaluation, which is based on the primary survey results, cannot be expected.

3) Methods of Analysis

In order to clarify the judgment standards for evaluations, quantitative analysis should be used as much as possible. To do this, it is necessary to clearly define the purpose and goal, as well as their indicators, at the start of a project. Along with describing what the projects purposes achieved and a evaluation by the Five Evaluation Criteria, evaluation using a five-level rating system is also possible. It is also necessary to fully analyze impeding factors.

4) Report Writing

An evaluation report should be easy to understand. When describing the implementation process, be sure to describe both its positive and negative aspects. When doing a questionnaire survey, be sure to attach the question items and their response to the report. It is best to describe results and evaluations basically in line with the project purpose and the indicators and activities of the PDM. The drawback to describing all the study results using only tables is that although the relationship between them is easy to understand, small letters and a lack of graphs and charts make it hard to read and understand. It is recommended to describe the most important findings in the main text. It is also necessary to make and present lists and charts of statistics and study results concerning these findings in the main text so that readers easily understand the reasons for the evaluation's conclusions.

(3) Purpose of a Secondary Evaluation

1) A Guidebook for Evaluations

There are various guidelines and manuals for evaluations. But they only describe basic theory and rules of thumb and often lack concreteness and clarity. If a secondary evaluation is used to present a high quality evaluation report objectively, these reports can then be used as a model for future reports. By producing reports like these over a period of several years, model reports will be available for every sector and cooperation scheme. As long as future evaluators prepare their reports according to the method and content of the models, report quality will be assured.

Among the 40 evaluation reports covered by this secondary evaluation, evaluation reports scoring high on 27 evaluation items include: the "Research and Development Project on High Productivity Rice Technology (in the Philippines)", the"Quality Improvement of Foundry Technology in Small and Medium Scale Industry (in Brazil)" project. These reports are also highly evaluated on project performance based on information given in the reports. Conversely, as for "The National Center for Environment (in Chile)" project, the evaluation report was evaluated as highly as the above reports, however, the evaluation of project performance was lower than average. Generally, there is a strong correlation between quality evaluation results (especially "Information Processing and Analysis" and "Usefulness" of "Five Evaluation Criteria") and evaluation results based on reports, but some reports of high quality do not highly evaluate the projects themselves, as in the case of Chile.

2) Impartiality of Evaluation Results

In JICA, the departments in charge of project management do a terminal evaluation. The evaluation by the department with full knowledge of a project is important, but the secondary evaluation by external experts assures that the transparency and objectivity of evaluations increases. Through these activities, the quality of terminal evaluations improves.

However, the same can be said for the external evaluation by experts. The evaluation results are objective in the sense that they are done by external experts who do not have a personal interest in the project. But it is a different issue whether the content of these evaluations is credible. As shown in this report, experts have biases when doing an evaluation. Basically, it is best to ask

multiple experts to evaluate the same project and clarify the evaluation categories in which they agree and disagree. This is not practical, however, from the perspective of time and cost.

But, this does not mean accepting the evaluation results without question. It is necessary to derive general knowledge, independent of a person's evaluation bias, including the experts' evaluation results, through the input of several other experts.

3) Developing a Feedback Mechanism on Evaluation Results

One reason why evaluation results are not read and utilized is that by the time of the evaluation the project is over and its results are too late to improve the project. Therefore, it is assumed that the results will be applied to similar projects in the future, but this requires generalizing the lessons and recommendations.

If many evaluations are conducted and multiple lessons and recommendations presented, there is the possibility that they may contradict one another. When this happens, it is necessary to generalize and authorize individual evaluation results between the time they are created and put to use. Furthermore, it is important to have an organizational framework that can act on the authorized recommendations. It is also necessary to verify that the recommendations are actually applied. In the past, in addition to problems in the content of evaluation results and their presentation, underutilized evaluations were the result of an insufficient organizational framework. This led to relying on common sense and the personal effort of those concerned to improve JICA projects.

A framework must confirm the lessons and recommendations of an evaluation by applying them and not leave recommendations as they are. The Advisory Committee on Evaluation hopes to play a part in this process.